Investigating clozapine health outcomes using electronic health records

Govind, Risha

Awarding institution:
King's College London

Investigating clozapine health outcomes using electronic health records

Risha Govind

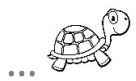
Thesis submitted for the degree of Doctor of Philosophy

Institute of Psychiatry, Psychology & Neuroscience

King's College London

2022

little by little, one travels far



ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisors. To Professor Cathryn Lewis, thank you for your unwavering guidance and giving me countless chances throughout my PhD, and for having faith in my ability to get the job done. To Professor Richard Dobson and Professor James MacCabe, thank you for always being there and guiding me at every step of the way. Thank you for allowing me to think out loud and always leading with kindness, patience, and compassion.

Next, I would like to thank my husband. My PhD journey was hard on you too, I am grateful that we stayed true. Thank you for being strong and resilient, especially on days I couldn't be. Your love, support, and companionship during the work-from-home years of the horrid pandemic truly helped me finish my PhD.

This PhD would not have been possible or as enjoyable without my two champions, Erica Mazaika and Paul Barton. I can never thank you enough for all your support and encouragement throughout this journey. I will forever be grateful for those long phone calls we had about resilience, self-compassion, courage, inner critic and inner child in times when I needed them the most. I would also like to thank the support services at King's College London, especially my academic mentor, Geoffrey Cantor, for mental health and wellbeing support and for introducing me to interesting ideas like 'individualisation' and self-compassionate habits like 'remembering to congratulate myself'.

Halfway through my PhD, the COVID-19 pandemic struck and changed the ways in which we were allowed to socialize. Doing a PhD in normal times is already so hard and isolating. I would like to thank the following authors for their audiobooks which kept me company whenever I felt isolated – Thich Nhat Hanh, Erling Kagge, Michael Singer, Eckhart Tolle, Oliver Sacks, Steve Peters, Andy Puddicombe, Amy Cuddy and Oprah Winfrey.

I would like to thank all my colleagues in the SGDP and the SLaM BRC Nucleus for supporting my research. In particular, I would like to thank Dr Daniela Fonseca de Freitas, Dr Richard Hayes, Dr Ehtesham Iqbal (Shami) and Megan Pritchard for guiding and supporting me in my research.

Last but not the least, I would like to thank all my friends and families for their support and kindness.

ABSTRACT

Background

Clozapine is the only evidence-based medication for treatment-resistant schizophrenia. However, clozapine is severely under-prescribed mainly because of clozapine-induced agranulocytosis, an adverse drug reaction of clozapine that occurs in 0.4% of clozapine-treated patients. Since there are no clinical predictors for clozapine-induced agranulocytosis, all clozapine users are required to be under strict blood test monitoring throughout the duration of their clozapine treatment.

Aims of this thesis

To use data from electronic health records to test the following hypothesis:

(i) Predictors for clozapine-induced agranulocytosis can be investigated using the results of the analysed electronics health records data published in (Iqbal *et al.*, 2020)

(ii) The frequency density of clozapine blood test results that indicate clozapine-induced agranulocytosis risk changes with clozapine treatment time

(iii) Clozapine-treated patients are at an increased risk of COVID-19 infection

(iv) Clozapine-treated patients are at an increased risk of severe outcomes of COVID-19 infection

Methods

All data were extracted from Clinical Record Interactive Search (CRIS), the de-identified electronic health records of South London and Maudsley NHS Foundation Trust (SLAM). All data extraction was performed using SQL. All analysis was performed using either R, python or STATA. The statistical methods used in this thesis are logistic regression, survival analysis and Cox proportional hazard models.

Results

We found that the data from (Iqbal *et al.*, 2020) was not informative for investigating predictors for clozapine-induced agranulocytosis, thus this study did not bear any significant results. However, it helped us to realise that the next step was to study the patterns of clozapine blood monitoring data.

From studying the patterns in clozapine blood monitoring results, we showed that the highest risk of clozapine-induced agranulocytosis is in the early months of treatments. The Kaplan-Meier survival curve and the incidence rates analysis showed that 75% of blood test results that indicated clozapine-induced agranulocytosis risk occurred within the first 6 months of clozapine treatment.

At the onset of the COVID-19 pandemic, we investigated the associations between clozapine treatment and increased risk of COVID-19. We found that clozapine-treated patients had an increased risk of COVID-19 compared with those who were on other antipsychotic medication (unadjusted HR = 2.62, 95% CI 1.73 - 3.96), which was attenuated after adjusting for potential confounders, including clinical contact (adjusted HR=1.76, 95% CI 1.14 - 2.72).

We followed up on the previous study to investigate the associations between clozapine treatment and increased risk of severe outcomes of COVID-19, namely COVID-related hospitalisation, COVID-related intensive care treatment, and death. We found that even though clozapine treatment appears to increase the risk of COVID-19 infection, it does not increase the risk of the severe outcomes of COVID-19.

Conclusion

In conclusion, electronic health records are a valuable resource for studying clozapine health outcomes. In particular, the CRIS data is a very informative resource for answering research questions related to mental health disorders.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADR	Adverse Drug Reaction
ANC	Absolute Neutrophil Count
BMI	Body Mass Index
BRC	Biomedical Research Centre
CLDS	Clinical Data Linkage Service
CLZ-ADR	Clozapine-induced Adverse Drug Reaction
CNRD	Central Non-Rechallenge Database
CRIS	Clinical Record Interactive Search
EHR	Electronic Health Records
ICD-10	Tenth Revision of the International Classification of Diseases
NLP	Natural Language Processing
PHM	Population Health Management
PLT	Platelets
SLAM	South London and Maudsley NHS Foundation Trust
VIEWER	Visualisation & Interaction With Electronic Records
WBC	White Blood Cells
ZTAS	Zaponex Treatment Access System

CHAPTER 1

1 INTRODUCTION

In this chapter, I provide the general background for my thesis. The chapter starts with a brief introduction on electronic health records and its usefulness in research. This is followed by an introduction of clozapine. I conclude this chapter with the motivation and outline of the thesis.

1.1 ELECTRONIC HEALTH RECORDS (EHR)

Electronic Health Records (EHR) are the collection of all documents that result from all interactions a patient has with the healthcare providers during their patient care (Ambinder, 2005). EHR includes data on the patient's medical history, diagnoses, current medications, past medications, demographics, progress notes, nurse's notes, clinician's notes, laboratory test results, medication prescriptions, medical images, incoming and outgoing correspondence with other healthcare providers (Xiao, Choi and Sun, 2018).

In addition to providing an essential repository for clinical records, the EHR forms a valuable resource for research. Unfortunately, the current computational solutions are not adequate to utilize EHR to its fullest potential in research studies. This is because just like the traditional paper-based health records, the EHR are also largely in a free-text clinical narrative format which has very little standardisation (Ross, Wei and Ohno-Machado, 2014). Limitations such as lack of standards in writing clinical notes, lack of structured data formats for patients' clinical information, incompleteness, and inaccuracy in EHR continue to lead

way for new computational solutions adapted for clinical data, for example, Natural Language Processing (NLP).

NLP techniques are used to extract information from large quantities of free-text clinical narratives to generate datasets that can be used for analysis (Jackson *et al.*, 2017). NLP algorithms are more sophisticated than basic keyword searches because they also assess the linguistic context around terms of interest, for example, temporal modifiers (e.g., "is currently on clozapine" versus "was previously on clozapine"). NLP techniques can therefore be used to turn text narrative into structured data to be used in research studies.

A typical EHR of a hospital is a big data source with millions of data points stored in hundreds of database tables. The information in EHR holds the potential to answer a wide range of research questions. However, to take advantage of this big data source, one needs to solve major challenges that currently exist with working with different aspects of EHR data. Researchers need to design their research studies such that they take the challenges into account. For example, EHR data hold the potential to thoroughly study the health outcomes of medications. This can be studied by using NLP tools to extract medication start and stop dates written in the clinical notes of the EHR, and then also extract all the potential covariates that were concurrent with the time the patient was on-treatment. Unfortunately, retrieving meaningful information on medication start and stop dates from clinical notes has a temporality problem, a major ongoing challenge of using EHR that remains unsolved (Cheng *et al.*, 2016). Temporality, which originated from the word "temporal", refers to the time dimension of the data, and in this context, it refers to identifying when the patient started and stopped taking medication. The temporality problem of EHR text-mining arises from the fact that there are countless ways in which clinicians write about a patient being

on medications in the free-text fields. This includes not writing anything about the medication in the clinical notes, which could imply that the clinician did not find it important to mention ongoing treatments that are effective and need no changes. For this reason, no mention of medications in clinical notes cannot be interpreted as an indication of the patient not taking the medication. Apart from the temporality problem, challenges to investigating medications in EHR data is also due to different clinicians having different writing styles, including their own abbreviations of medication names. This lack of standardisation in writing in the clinical notes that a patient is on a medication is a major computational challenge for mining medication data from EHR.

Various research efforts are underway to address the several unsolved problems of the EHR and progress is happening steadily, even though slowly (Zhao and Henriksson, 2016; Che *et al.*, 2018; Bagattini *et al.*, 2019; Chen *et al.*, 2021; Pedersen *et al.*, 2021).

1.1.1 Using EHR for research

EHR is a rich real-world dataset resource containing all patient data that can provide a major platform for performing retrospective and observational studies. The longitudinal real-time health records data stored in EHR make them an exceptional source to construct phenotypes of patients. Recently, substantial progress has been made towards making all patient's records electronic and paperless, including by the UK government (Crane and Bunn, 2016).

In recent years, considerable effort has been made to help researchers access and analyse electronic health records, for example, de-identifying patient records so databases can be linked and made accessible to researchers to perform secondary analysis on the data

sources from multiple service levels. Examples of some centralised databases available for research are the Hospital Episode Statistics (HES) database that provides secondary care admission information; the Office for National Statistics (ONS) database that provides various levels of data such as mortality information and Indices of Multiple Deprivation (IMD) that scores the deprivation levels of neighbourhoods.

There has also been an increase in re-developing case registers which once were only for the purpose of patient healthcare management with limited research focused mainly on outcomes and service evaluations (Bloor, 1995).

The South London and Maudsley NHS Foundation Trust (SLAM) is one of the world's largest mental health research institutions. SLAM's Electronic Health Records (EHR) started in 2006. Two years later, SLAM developed a Clinical Record Interactive Search (CRIS) system to create a platform with a fully de-identified case register so researchers can perform secondary analysis within robust data security and governance framework (Stewart *et al.*, 2009). Today, the SLAM has the largest mental health case register in Europe, the Biomedical Research Centre (BRC) SLAM case register.

The BRC SLAM case register not only contains the de-identified EHR of the SLAM patients but also has (1) data linkage to external databases such as HES and ONS, and (2) results from in-house custom built NLP apps that were developed by designated specialised NLP data scientists recruited to build and maintain the apps. The NLP team work closely with researchers to make sure their priorities are in harmony. Some examples of the NLP apps developed by the BRC SLAM team are the medication app which gives the start and end dates of medications; the diagnosis app provides information of the confirmed diagnoses

with the date of diagnosis; the Body Mass Index (BMI) apps provide information on the BMI values and with the record date; the inpatient app provides information on the dates each patient was hospitalised.

Some examples of EHR research that have shed insights are the early detections of risks of suicide (Jayasinghe *et al.*, 2020), enhancing detection and prediction of diseases (Irving *et al.*, 2021) and revealing information on adverse drug reactions (ADR) (Iqbal *et al.*, 2020).

The aim of the ADR of medications (Iqbal *et al.*, 2020) study was to develop an algorithm for identifying ADR-related clinical events from free-text clinical notes of EHR data and share this information with clinicians to help them improve their understanding of ADRs. The algorithm was developed using EHR data from three separate mental health trusts in the UK, which comprised over 50 million documents from over 500,000 patients. The three mental health trusts were SLAM, the Camden and Islington NHS Foundation Trust and the Oxford Health NHS Foundation Trust. Similar to SLAM, the other two trusts also had the CRIS system. All EHR data were accessed by the CRIS platform, thus all data used in the study was fully de-identified. The algorithm extracted the following three pieces of information from the free-text fields: (1) list of patients who were on clozapine treatment (2) clozapine treatment start and stop dates (3) ADRs that occurred during clozapine treatments. The algorithm identified 2,835 clozapine-treated patients across the three mental health trusts. From the results of the algorithm, it was found that hospital admission showed a significant association in 30 out of the 33 ADRs that the algorithm identified. This paper sheds insights on the importance of ADR extraction from the free-text fields of the EHR, as it can help clinicians and researchers better understand ADRs. This is particularly important for medications like clozapine, which is underutilized because of its ADR.

1.2 CLOZAPINE TREATMENT

Clozapine is an antipsychotic, and the only evidence-based treatment for schizophrenia patients who do not respond to, or cannot tolerate the conventional antipsychotic drugs (Wahlbeck *et al.*, 1999; Chakos *et al.*, 2001). Approximately 23% of schizophrenia patients fall in this category. Despite clozapine currently being the gold standard treatment for them, it is estimated that only 5% to 20% of clozapine-eligible patients receive clozapine treatment (Meltzer, 2012; Olfson *et al.*, 2016; Siskind *et al.*, 2021)

1.2.1 Schizophrenia

Schizophrenia is a devastating and severe progressive psychiatric disorder and it affects approximately 1% of the global population (Rössler *et al.*, 2005). Schizophrenia can be debilitating and chronic, and is amongst the world's top 15 leading causes of long-term disability (GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, 2018).

The symptoms of schizophrenia include hallucinations, delusions, cognitive impairment, thought disorder (disorganized ways of thinking), abnormal behaviour and disorganized speech. Schizophrenia symptoms are typically described using these three broad categories: positive symptoms, negative symptoms, and cognitive impairment (Liddle, 1987). Positive symptoms characterise the changes in thoughts and behaviour that are "added on" to a person's experiences such as hallucinations and delusions. Negative symptoms characterise the deficit states of a person's experiences such as the inability to show emotions, apathy, and disorganized speech (Dollfus and Lyne, 2017). Cognitive impairment refers to difficulties in attention, concentration, and memory. The current understanding of the aetiology of schizophrenia is that it cannot be narrowed down to one single factor but to a complex

interplay between genetic and environmental risk factors (Weinberger and Harrison, 2011; Howes and Murray, 2014).

The diagnosis of schizophrenia in the UK and most of Europe is currently based on the Tenth Revision of the International Classification of Diseases (ICD-10) guidelines which are set by the World Health Organization (WHO, 1992). The ICD-10 codes for schizophrenia and schizophrenia-spectrum disorders are from F20 to F29. This block includes schizophrenia, schizotypal disorder, persistent delusional disorders, schizoaffective disorders, and a larger group of acute and transient psychotic disorders.

Once schizophrenia is diagnosed, the severity of schizophrenia symptoms can be measured using a standardised rating scale. Today, there are several well-established rating scales that are used in the clinic to track the changes in symptoms over time. These ratings scales assesses the positive and negative schizophrenia symptoms and the common ones are the Positive And Negative Syndrome Scale (PANSS) (Kay, Fiszbein and Opler, 1987), the Brief Psychiatric Rating Scale (BPRS) (Overall and Gorham, 1962), the Scale for the Assessment of Negative Symptoms (SANS) (Andreasen, 1989) and the Scale for the Assessment of Positive Symptoms (SAPS) (Andreasen, 1984).

1.2.2 Pharmacological treatment of schizophrenia

Antipsychotics are the cornerstone medication for the treatment of schizophrenia. Antipsychotic medications reduce up to 60% of symptoms in first-episode schizophrenia patients by alleviating most of the positive symptoms, thereby significantly reducing the relapse of the disease (Kahn *et al.*, 2008; Leucht *et al.*, 2012). There are two categories of

antipsychotic medications, the first-generation antipsychotics, and the second-generation antipsychotics.

The first-generation antipsychotics, also referred to as 'typical' or 'conventional' antipsychotics, were the first antipsychotics to be developed. The introduction of chlorpromazine in the 1950s revolutionized the treatment of schizophrenia, this was the first first-generation antipsychotic (López-Muñoz *et al.*, 2005). Examples of first-generation antipsychotics that are currently licensed for use in the UK are chlorpromazine, flupentixol, haloperidol, levomepromazine, pericyazine, perphenazine, pimozide, prochlorperazine, promazine, sulpiride, trifluperazine, zuclopenthixol (MHRA, 2005; BNF, 2020).

The second-generation antipsychotics, also referred to as 'atypical' antipsychotics were first developed in the 1990s. The 8 second-generation antipsychotic drugs that are currently licensed for use in the UK are amisulpride, aripiprazole, clozapine, lurasidone, olanzapine, paliperidone, quetiapine, risperidone (MHRA, 2005; BNF, 2020).

The leading motivation for the development of the second-generation antipsychotics is also the main difference between the two categories of antipsychotics. The first-generation antipsychotics have an increased burden of the debilitating extrapyramidal side effects (Kane *et al.*, 1988). Extrapyramidal side effects refer to the involuntary movement disorders such as tremors, spasms, repetitive movements, dry mouth, and behaviour that mimics Parkinson's disease. The second-generation antipsychotics have a lower risk of the extrapyramidal side effects. However, they do have an increased burden of other side effects like weight gain and metabolic side effects.

Given the large number of antipsychotics available, none of them being without side effects, a significant obstacle in treating schizophrenia is to balance the management of schizophrenia symptoms in parallel with managing the side-effects from the antipsychotic medications. The clinical decision to prescribe involves the trial-and-error process of weighing the side effects profile of each drug against each drug's efficacy profile towards the symptoms the individual patient is experiencing (Hamann *et al.*, 2005; Taylor, Barnes and Young, 2018). The patient's past treatment response, preferred route of administration and co-morbidities are also considered when making this decision.

Once an antipsychotic medication is selected and prescribed, then it is titrated to its minimum effective dose. Drug titration is the process of adjusting the dose of a medication to achieve the best clinical response of the drug with minimum side effects. The patient is closely monitored for the therapeutic effects as well as any side effects and compliance. Compliance refers to the patient's act of taking the medication as prescribed and on schedule. If the compliance is a major issue, then depot antipsychotics are considered. Depot antipsychotics are administered via injections and the medication slowly releases into the body over several weeks (Ting *et al.*, 2019). If a patient develops severe side effects to the medication or persistently finds it ineffective, then the patient is tried on a different antipsychotic until an effective antipsychotic is found (Stroup and Gray, 2018).

Unfortunately, a significant number of patients find many of the antipsychotic medications ineffective. Approximately 23% of schizophrenia patients continue to experience the debilitating symptoms despite trying different antipsychotic medications (Lindenmayer, 2000; Mortimer *et al.*, 2010; Siskind *et al.*, 2021).

1.2.3 Treatment-resistant schizophrenia

When a patient's symptoms persist after trying two or more different antipsychotic medications at adequate therapeutic doses and for an adequate duration, they are considered to have treatment-resistant schizophrenia (TRS) (Lehman *et al.*, 2004; Howes *et al.*, 2017).

Beyond the above definition of TRS, existing guidelines vary in the stringency of details of TRS definition. For example, the NICE guidelines have no recommendations on the adequate treatment duration of the two failed antipsychotic medications whereas the Maudsley prescribing guidelines does (National Institute For Clinical Excellence, 2014; Taylor, Barnes and Young, 2018). To be considered TRS, the Maudsley prescribing guidelines requires a minimum of 3 weeks treatment duration for the first failed antipsychotic followed by a minimum of 6 weeks treatment duration for the second failed antipsychotic. The NICE guideline also requires the 2 failed antipsychotic trials to be sequential whereas the Maudsley prescribing guidelines does not have this requirement.

The first rigorous criteria to define TRS was presented by Kane et al in 1988 (Kane *et al.*, 1988). The criteria were based on the persistence of positive symptoms despite the patient receiving 'adequate treatment'. As new evidence emerged, the criteria evolved with time, from the initial rigorous criteria to a broader range of definition for TRS. Today, various definitions and criteria exists that define TRS as a clinical entity. In 2017, the Treatment Response and Resistance in Psychosis (TRRIP) working group studied the various existing criteria to define TRS and the variations between the different guidelines and published consensus criteria to standardise the assessment and definition of TRS (Howes *et al.*, 2017).

The TRIPP working group's minimum consensus criteria for assessing TRS are (Howes *et al.*, 2017):

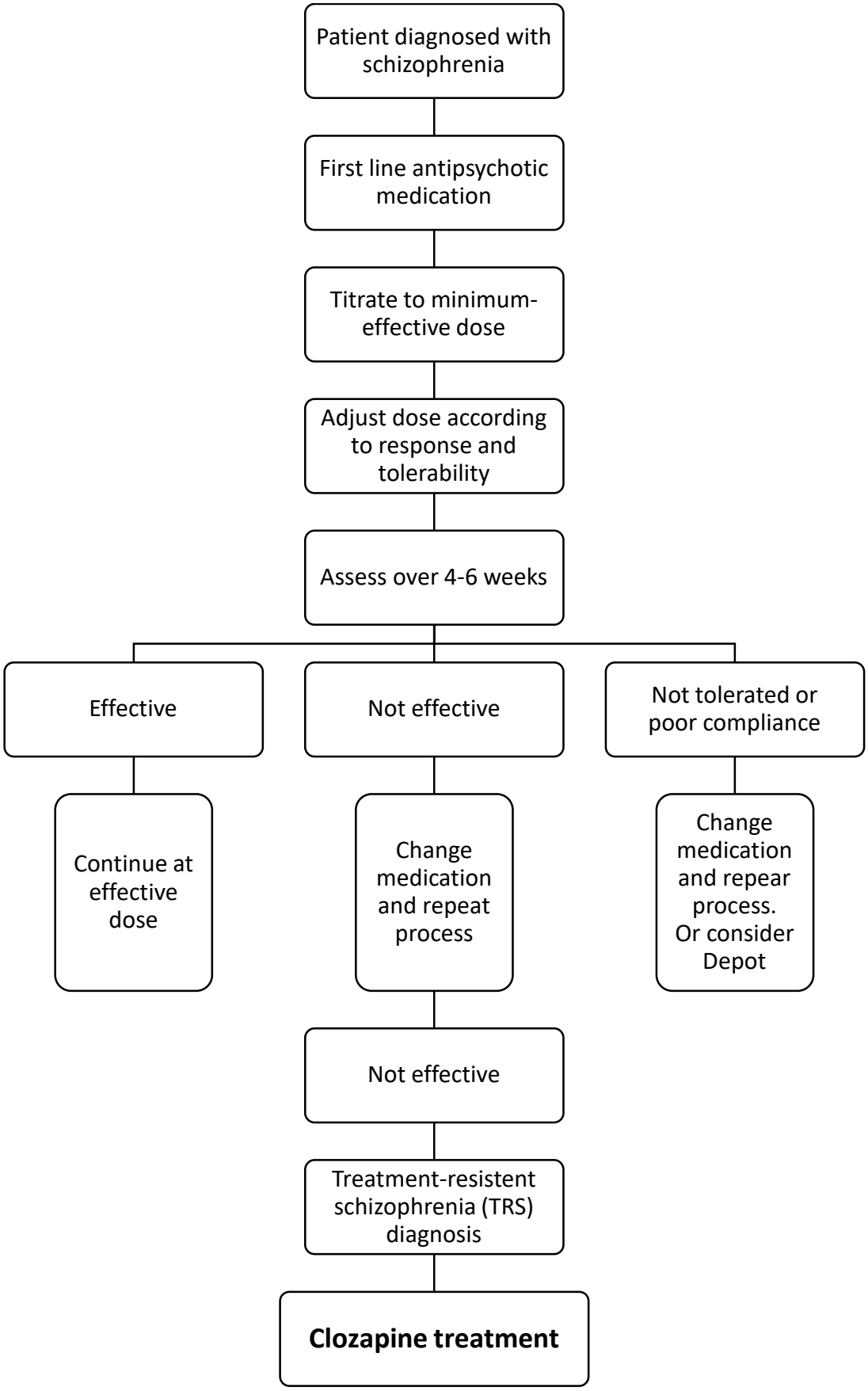
- Current symptoms must be assessed using the standardised rating scales (e.g., PANSS, BPRS, SANS, SAPS)
- Current symptoms must be of at least moderate severity for a minimum duration of 12 weeks.
- The information for assessment of response to the failed antipsychotic treatments must be gathered from patients, carers, hospital staff, case notes, pharmacy dispensary information
- There must be at least 2 different failed antipsychotic medications, each with a minimum treatment duration of 6 weeks and with a minimum therapeutic dose that is equivalent to 600mg of chlorpromazine equivalent per day. The minimum and mean statistics of the treatment duration and medication dose must be recorded.
- The assessment for medication compliance must come from at least 2 different sources, such as from pharmacy information and clinical notes. Additionally, at least one blood test for monitoring the antipsychotic serum levels in the blood must be available.

Even though there is still no universal agreement on a robust set of criteria for defining TRS, the consensus criteria outlined by the TRIPP working group highlight the 3 key concepts that are common between the existing clinical guidelines: 1) a confirmed schizophrenia diagnosis through a recognised rating scale, 2) two failed treatments of adequate dose and duration, and 3) persistence of significant symptoms despite adequate pharmacological treatments.

Despite the differences in the guidelines when it comes to defining TRS, all guidelines agree that the TRS are amongst the most severely distressed and disabled of all mental health patients. TRS is a serious problem because it involves severely ill patients who are not on an effective medication. As a result, compared to the average schizophrenia population, TRS patients visit the hospital more and require longer hospitalisation (Lindenmayer, 2000). Also, TRS patients are at a higher risk for suicide, have a lower quality of life, are more likely to have substance abuse problems and therefore, require more social and healthcare (Kennedy *et al.*, 2014).

Fortunately, there is one clinically proven drug that is highly effective on TRS patients and proven to reduce their symptoms and reduce hospitalization (Wahlbeck *et al.*, 1999; Stroup *et al.*, 2016). This drug is clozapine. However, there is a long pharmacological process required to qualify for clozapine treatment, as illustrated in Figure 1.a.

Figure 1.a: Pharmacological treatment of schizophrenia and TRS (NICE, 2014)



1.2.4 A brief history of clozapine

In 1959, clozapine was first synthesised (Hippius, 1999). This was just a few years after the discovery of the first antipsychotic medication, which is still considered the biggest revolution in the treatment of schizophrenia (Hippius, 1999; Seeman, 2014). All antipsychotic medications that were discovered prior to clozapine were burdened with extrapyramidal side-effects (Kane *et al.*, 1988). Clozapine was the first fully effective antipsychotic medication with the unique property of rarely causing extrapyramidal side-effects (Stephens, 1990). In those days, there was a really strong dogma in the field of psychiatry that extrapyramidal symptoms were a *conditio sine qua non* (i.e. an essential condition) for antipsychotic medications to be fully effective (Van Rossum *et al.*, 1970). Hanns Hippius, a German psychiatrist who was part of the team that investigated clozapine described the traditional thinking of the psychiatric community of the time as (Hippius, 1989):

> *… we discovered to our surprise that clozapine, in contrast to all other compounds, had no extrapyramidal effects despite being a fully effective antipsychotic. This finding was almost unbelievable, because at that time it was a part of psychopharmacological dogma that extrapyramidal effects went in tandem with antipsychotic efficacy.*

The first breakthrough in the development of treatment for schizophrenia was the discovery of the first antipsychotic medication. The psychiatric community welcomed this discovery with enthusiasm which resulted in a boost in research and further discoveries of similar antipsychotics, all of which were found to be associated with extrapyramidal side-effects. The association of antipsychotics with extrapyramidal side-effects became such a strong

dogma in the field that when the second breakthrough of the field occurred, the discovery of clozapine, an antipsychotic that is almost without extrapyramidal symptoms, the psychiatric community responded with suspicion and clozapine was not given serious consideration (Crilly, 2007). This limited interest in clozapine became a major challenge and slowed down clozapine research.

In 1971, clozapine became the first antipsychotic with practically no extrapyramidal side-effects to be introduced in clinical practice in Europe. It took over a decade for clozapine to be accepted by the psychiatric community and be recognised as a viable treatment for schizophrenia. The superior efficacy of clozapine compared to all other antipsychotic medications available were reported in controlled studies (Rodová *et al.*, 1973; Ekblom and Haggstrom, 1974; Gerlach *et al.*, 1974; Fischer Cornelssen and Ferner, 1976). However, this increasing momentum came to an abrupt halt when clozapine's association with severe blood dyscrasia was discovered (Hippius, 1999).

On 27th September 1975, the journal Lancet published a report from Finland that 16 of their clozapine treated patients developed severe blood dyscrasia within 4 months of starting clozapine treatment, 8 of whom had fatal outcomes (Idänpään-Heikkilä *et al.*, 1975). Following this, use of clozapine was dramatically reduced, the research on the drug was stopped, and clozapine was taken off the market (Khokhar *et al.*, 2018). Even before clozapine could be used extensively as a treatment for schizophrenia, it was essentially pulled from the market and its use was discontinued in clinical practice.

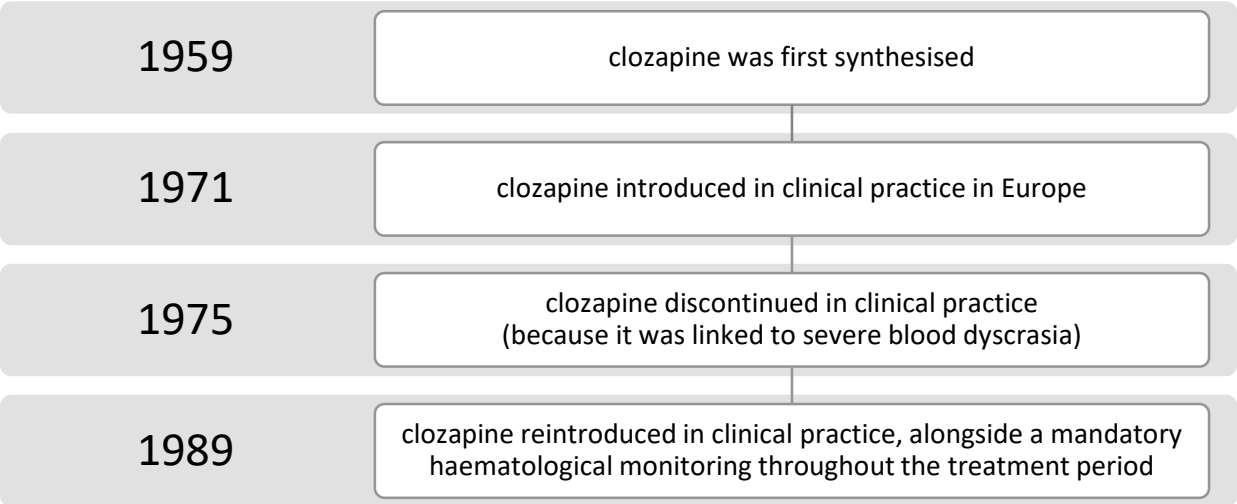
In 1988, a seminal study by John Kane and colleagues led to the reintroduction of clozapine (Kane *et al.*, 1988). The study demonstrated the high efficacy of clozapine in treating the

schizophrenia patients who were refractory to all other exciting antipsychotic medications. The study also demonstrated that a safe way to use clozapine and prevent clozapine-related severe blood dyscrasia is via careful, regular monitoring of blood cell counts.

In 1989, clozapine was re-introduced with mandatory routine haematological monitoring throughout the duration of the clozapine treatment. In addition, strict guidelines need to be followed before clozapine can be prescribed.

The re-introduction of clozapine significantly changed the treatment approach for patients with treatment-resistant schizophrenia (TRS). With emerging research, clozapine is now considered the most effective antipsychotic medication for the treatment of TRS. Its efficacy is so superior that 60% - 70% of users who don't respond to any other antipsychotic medication, respond to clozapine (Meltzer, 1992). The timeline from the discovery to the present day clinical use of clozapine is summarised in Figure 1.b.

Figure 1.b: The timeline of clozapine - from discovery to present day use in clinical practice



1.2.5 Strict guidelines to prescribing clozapine in the UK

In the UK, clozapine use is only licensed for patients with TRS. Before a TRS patient can be prescribed clozapine, there are the strict clozapine initiation guidelines that need to be followed (BNF, 2020).

First, the patient must be registered with a clozapine monitoring service, which is provided by the manufacturers. There are 3 clozapine monitoring services in the UK: Zaponex Treatment Access System (ZTAS), Clozaril Patient Monitoring Service (CPMS) and Denzapine Monitoring Service (DMS). Before a patient can be registered with a monitoring service, the consultant responsible for prescribing and treating the patient with clozapine and the designated pharmacist responsible for dispensing the clozapine to the patient must be registered with the same monitoring service.

Regardless of which monitoring service a patient is registered to, if sign of blood dyscrasia is seen in a patient's clozapine blood monitoring, this information is recorded in the Central Non-Rechallenge Database (CNRD) which is shared across all monitoring services. Before a patient can be registered with a monitoring service, they need to pass the CNRD check.

Once the patient is registered with a monitoring service, a baseline blood test is performed. This is to check the blood count values before the initiation of clozapine treatment. The blood test results affect the decision on whether clozapine can be prescribed to this patient or not. At the discretion of the monitoring service and the treating consultant, baseline test can be repeated until a valid blood test result is observed.

If a valid blood test is recorded and all the due diligence is successfully completed by the monitoring service for registering the patient, then the monitoring service informs the treating consultant that they are now allowed to prescribe the clozapine for 7 days at a time

for the next 18 weeks or until the blood monitoring results indicate a red alert. Also, the monitoring service informs the designated pharmacist who is registered to be responsible to prescribe clozapine to the patient that they can now prepare the first dispense of clozapine. The flowchart of the clozapine monitoring process is shown in Figure 1.c.

1.2.6 Mandatory haematological monitoring of clozapine patients in the UK

Clozapine is recognized as the gold standard treatment for TRS (McEvoy *et al.*, 2006; National Institute For Clinical Excellence, 2014). However, clozapine is linked to several side effect such as weight gain (Marteene *et al.*, 2019), hypersalivation (Chen *et al.*, 2019), constipation (Every-Palmer *et al.*, 2016) and postural hypotension (Nielsen *et al.*, 2011). Clozapine is also linked to adverse effects such as myocarditis (Siskind, Sidhu, *et al.*, 2020), thromboembolism (Hägg, Spigset and Söderström, 2000) and hematologic dyscrasia (Meyer *et al.*, 2015). It is the haematological adverse effects of clozapine that warrants the mandatory routine haematological monitoring.

The haematological adverse effects of clozapine are neutropenia and agranulocytosis. Agranulocytosis and neutropenia are blood dyscrasia characterised by a severe reduction in neutrophils in the blood. Neutrophils are a type of white blood cells and severe reduction in neutrophil can be life-threatening. The normal range for neutrophil count is between 2 x 10^9/L and 7.5 x 10^9/L. Neutropenia occurs when the neutrophil count drops below 1.5 x 10^9/L. Agranulocytosis, which is also known as severe neutropenia, occurs when the neutrophil count drops below 0.5 x 10^9/L. In clozapine-treated patients, the prevalence of neutropenia and agranulocytosis is reported to be 3.8% and 0.4%, respectively (Myles *et al.*,

2018; Li *et al.*, 2020). Stringent haematological monitoring is used to detect early sign of neutropenia. The clozapine treatment is stopped at the first sign of neutropenia.

In the UK, white blood cell count (WBC) and absolute neutrophil count (ANC) monitoring is a requirement for continuing clozapine therapy. All patients on clozapine must undergo at least weekly blood monitoring for the first 18 weeks, at least fortnightly from 19 to 52 weeks, and then at least four-weekly for the rest of their clozapine treatment. The frequency of monitoring is increased if the patient develops symptoms of possible infections, for example, flu-like symptoms or fever. The increased monitoring persists until the symptoms subside. Drops or downward trends in the WBC/ANC can also lead to increased monitoring.

The blood monitoring results are reported back in a traffic light color-coded system: green, amber and red. Table 1.a shows the WBC and ANC counts needed for each colour code. Further patient management is based on the monitoring results. The clozapine treatment is continued with no further action required after a green alert. If there is an amber alert in the monitoring, then monitoring is changed to twice weekly until the results are in the green alert range. If there is a red alert in the monitoring, which is the patient's WBC is less than 3.0×10^9/L and/or the neutrophil count is less than 1.5×10^9/L then the clozapine treatment is stopped immediately, all supply of clozapine is removed from the patient and an emergency blood test is arranged to confirm the red alert. If the red alert is confirmed, then the patient is entered onto a CNRD to ensure that they are never prescribed clozapine again, except under exceptional circumstances.

Table 1.a : UK clozapine monitoring categories and colour alerts for routine blood tests (ZTAS, 2018)

Colour alert	WBC $\times 10^9$/L	ANC $\times 10^9$/L	Guidance
Green	>3.5	>2.0	Continue treatment
Amber	3.0-3.5	1.5-2.0	Continue, but monitor twice weekly until alert turns green
Red	<3.0	<1.5	STOP CLOZAPINE TREATMENT IMMEDIATELY. Arrange emergency blood test to confirm the red alert.

The reference values for the colour-codes of the blood test are slightly different if the test is for a clozapine initiation (baseline test) or is after an interruption in clozapine treatment. This is shown in Table 1.b.

Table 1.b : UK clozapine monitoring categories for clozapine initiation baseline test (ZTAS, 2018)

Colour alert	WBC $\times 10^9$/L	ANC $\times 10^9$/L	Guidance
Green	>4.0	>2.5	Clozapine treatment can be initiated. Clozapine can be prescribed and dispensed, but only for 7 days at a time.
Intermediate Amber	3.5-4.0	2.0-2.5	Clozapine treatment may be initiated at the discretion of the treating consultant. Additional blood testing is advised.
Amber	3.0-3.5	1.5-2.0	Clozapine treatment **cannot** be initiated. Additional blood testing is required.
Red	<3.0	<1.5	Clozapine treatment **cannot** be initiated. The cause of the red alert must be investigated.

Any break in clozapine treatment that is for more than a week (for patients on weekly monitoring), or for more than 4 weeks (for patients on fortnightly or 4-weekly monitoring) is investigated by the monitoring service. If the monitoring service deems the treatment as "interrupted", then the patient will need to re-register and the reference values for their next blood test will be the same as the reference value for patients who are initiating clozapine treatment.

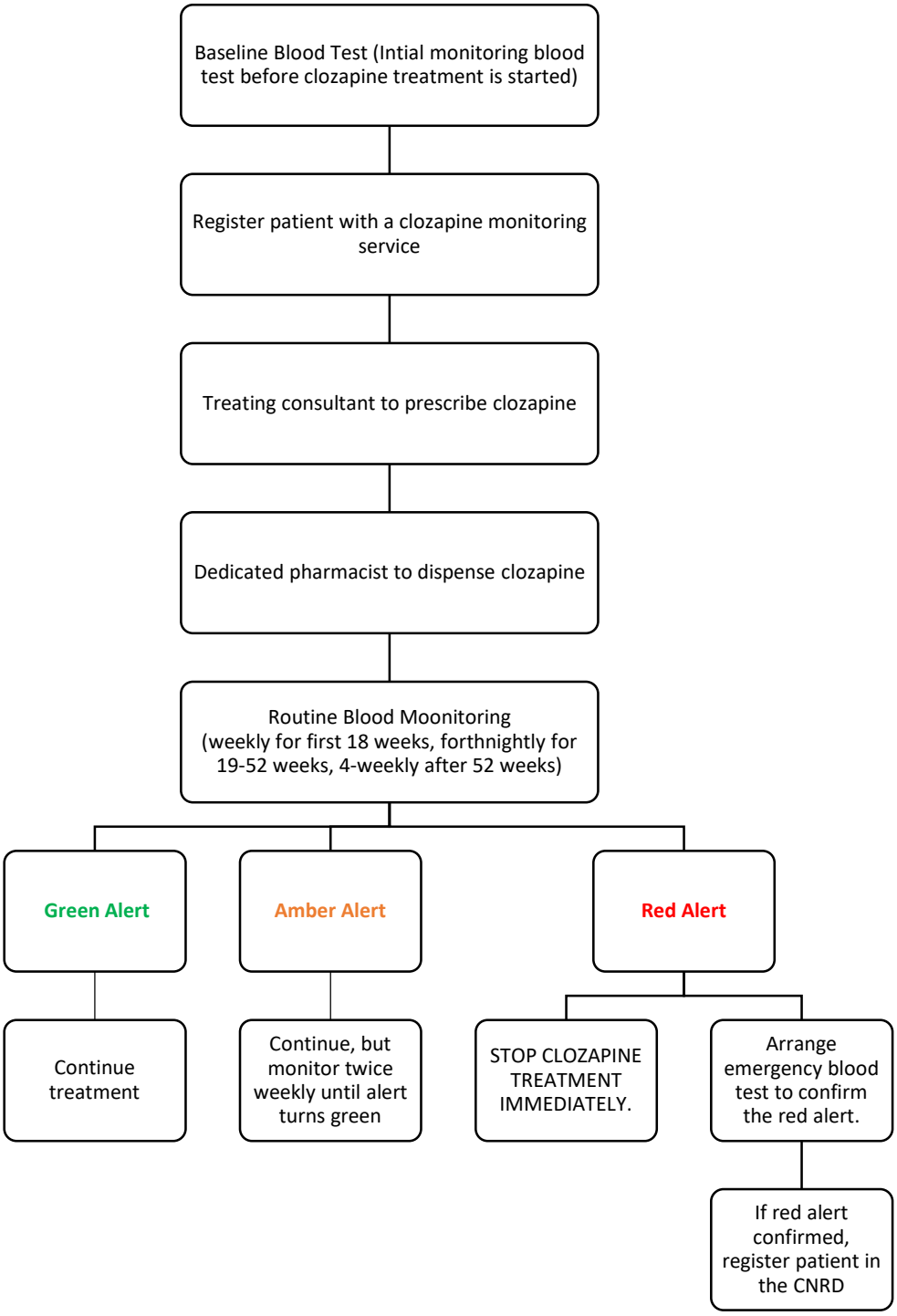
Benign Ethnic Neutropenia (BEN) is defined as the occurrence of neutropenia, which is defined by data from white population, in healthy individuals of non-white ethnicity (Haddy, Rana and Castro, 1999). BEN occurs in 25 – 50% of patients from a black ethnic background. (Shoenfeld *et al.*, 1988; Munro *et al.*, 1999; Rajagopal, 2005). For this reason, patients with black ethnic background are monitored based on a different set of reference values where lower thresholds of WBC and ANC are used before further monitoring or stopping treatment (Table 1.c).

Table 1.c : UK clozapine monitoring categories and colour alerts for patients with Benign Ethnic Neutropenia

Colour alert	WBC $\times 10^9$/L	ANC $\times 10^9$/L	Guidance
Green	>3.0	>1.5	Continue treatment
Amber	2.5-3.0	1.0-1.5	Continue, but monitor twice weekly until alert turns green
Red	<2.5	<1.0	STOP CLOZAPINE TREATMENT IMMEDIATELY. Arrange emergency blood test to confirm the red alert.

Under exceptional circumstances, patients who are registered in the CNRD can be rechallenged on clozapine. By July 2021, there were approximately 4000 patients registered on the CNRD in the UK, of whom 20% were estimated to have been rechallenged on clozapine (Oloyede *et al.*, 2021). The rechallenge process involves the treating consultant to formally confirm the awareness of the risks associated with rechallenging clozapine and the approval of the medical advisors at the monitoring service. Rechallenge clozapine treatments are classed “off-licence”.

Figure 1.c: Clozapine haematological monitoring flow chart



1.2.7 Underuse of clozapine

The mandatory haematological clinical monitoring and the strict initiation guidelines have proven to be highly effective in preventing fatal outcomes due to clozapine-induced neutropenia or agranulocytosis. To date, there have only been 8 fatalities from clozapine in the UK (MHRA, 2020). Unfortunately, the rigorous nature and the inconvenience of the monitoring is a major cause of clozapine underuse for TRS patients, despite clozapine being the only evidence-based effective treatment for them (Nielsen *et al.*, 2016).

Clozapine is the gold standard antipsychotic for TRS. If the patient's symptoms persist after trying two different antipsychotic medications at adequate therapeutic doses and for an adequate duration, then clozapine should be offered (Taylor *et al.*, 2012; National Institute For Clinical Excellence, 2014). However, this does not always happen. It is estimated that only 5% to 20% of clozapine-eligible patients receive clozapine treatment (Olfson *et al.*, 2016). The main reasons behind the underuse of clozapine are:

- The hesitation on the part of clinicians and/or patients due to fear of side effects.
- Complex pathway to qualify for clozapine use such as the treating consultant and a designated pharmacist need to be registered with a monitoring service before the patient can be registered with them. The process of registering a patient with a monitoring service is also complex.
- The difficulties with adhering to the strict and complex pathway to initiate clozapine use once the patient is successfully registered with the monitoring service.
- The difficulties with adhering to the rigorous intensive haematological monitoring throughout the clozapine treatment. For example, breaks in treatments are

investigated by the monitoring service and they could warrant to redo the initiation process if they deem the break in treatment was significantly long. Unfortunately, most clozapine patients suffer from very severe schizophrenia therefore breaks in their medication adherence are common.

- Patient's refusal due to dislike of phlebotomy, needle phobia or the inconvenience of the routine blood monitoring.
- Clinician unfamiliarity with the use of clozapine and the complex pathway to initiate clozapine use.

These reasons cause significant delays to early adoption of clozapine. They also become a barrier to using clozapine and lead to early discontinuation of clozapine treatment or the use of unjustified antipsychotic polypharmacy (Oloyede *et al.*, 2021). Antipsychotic polypharmacy refers to the co-prescription of more than one antipsychotic medication, and in this context, it refers to the co-prescription of more than one non-clozapine antipsychotics. Approximately 23% of schizophrenia patients do not respond to any antipsychotics other than clozapine (Lindenmayer, 2000; Mortimer *et al.*, 2010; Siskind *et al.*, 2021). For most of these patients, clozapine treatment is the only way for them to live outside the hospital. The antipsychotic polypharmacy can be harmful as it is associated with extrapyramidal side-effects, metabolic syndrome and higher incidence of hospitalisations (Correll *et al.*, 2007; Weinmann, Read and Aderhold, 2009; Barnes and Paton, 2011; Torniainen *et al.*, 2015; Kadra *et al.*, 2018).

1.3 THESIS MOTIVATION

Clozapine has the potential for wider use if the burden of the monitoring process can be reduced, and this could be transformational for many patients with TRS. The motivation of this thesis is to interrogate electronic health records (EHR) to improve our understanding of the underlying mechanisms of the effects of clozapine.

1.4 THESIS OUTLINE

This is a thesis incorporating three journal articles, one of which is published, one is in press, and one is a preprint, at the time of writing. These articles are incorporated in chapters 4, 5 and 6.

Chapter 2: This Methods chapter contains an in-depth description of data and the methodologies used in this thesis. The EHR from CRIS data and the tools used to extract information from it are described, in addition to the main exposure of interest, outcomes, covariates, and the main statistical analyses.

Chapter 3: This chapter is a demonstration of the challenges of working with EHR data for research. The study uses EHR from CRIS to investigate variables that are associated with clozapine-induced blood dyscrasia in SLAM patients. Unfortunately, some challenges of working with EHR data impacted this study to achieve its aim, and these are described in detail within the chapter. The skills learnt from this chapter were informative in designing the study of the other research chapters.

Chapter 4: This chapter integrates clozapine blood monitoring results data with EHR data from CRIS to define clozapine treatment periods, and then to identify incidence rates of

clozapine-induced blood dyscrasia against the length of clozapine treatment. Our findings showed a contrast between the relatively high density of blood dyscrasia incidences at the beginning of clozapine treatment which significantly reduces after 6 months of treatment which remained low thereafter. This chapter incorporates the preprint of the paper.

Chapter 5: This chapter uses data from CRIS to investigate whether clozapine treatment, compared to non-clozapine antipsychotic treatment, was associated with an increased risk of COVID-19 infection. Our findings suggested that receiving clozapine treatment is associated with increased COVID-19 risk, compared to receiving any other type of antipsychotic treatment. This chapter incorporates the paper the was published in May 2020 in British Journal of Psychiatry.

Chapter 6: This chapter incorporates the follow-up paper from Chapter 5. The CRIS data was used to investigate associations between clozapine treatment and increased risk of adverse outcomes of COVID-19, namely COVID-related hospitalisation, intensive care treatment, and death, among patients taking antipsychotics with schizophrenia-spectrum disorders. Our findings showed that even though we previously found an association between clozapine treatment and COVID-19 infection, we found no evidence that clozapine treatment puts patients at increased risk of adverse outcomes of COVID-19. This chapter incorporates the paper the was accepted for publication in European Neuropsychopharmacology in January 2022.

Chapter 7: This Discussion chapter summarises the key findings of the thesis. The strengths as well as the limitations of the work carried out are discussed, in addition to the implications of the findings and recommendations for the future research to follow.

1.5 CONTRIBUTION STATEMENTS

Chapter 3: I designed the study and performed all the data extraction and statistical analysis for this chapter.

Chapter 4: Amelia Jewell was responsible for creating the data linkage between the ZTAS data and the CRIS data. Upon the completion of the data linkage procedure, James MacCabe and I conceived and designed this study which is based on the linkage data. I performed all the data extraction and statistical analysis as well as wrote the manuscript. All authors critically reviewed the manuscript and approving the final version.

Chapter 5: This study was conceived by Richard Hayes and James MacCabe. They are both joint senior authors on the manuscript. Daniela Fonseca de Freitas, Richard Hayes, James MacCabe and I designed the study. With the support and guidance from other authors, I performed all the data extraction and statistical analysis for this chapter; Megan Pritchard supervised the data extraction part; Daniela Fonseca de Freitas supervised the statistical analysis part. I wrote the Methods and Results section of the manuscript. All authors contributed to manuscript preparation and approving the final version.

Chapter 6: This study was conceived by Richard Hayes and James MacCabe. They are both joint senior authors on the manuscript. Daniela Fonseca de Freitas, Richard Hayes, James MacCabe and I designed the study. With the support and guidance from other authors, I performed all the data extraction and statistical analysis for this chapter; Megan Pritchard supervised the data extraction part; Daniela Fonseca de Freitas supervised the statistical analysis part. I wrote the Methods and Results section of the manuscript. All authors contributed to manuscript preparation and approving the final version.

CHAPTER 2

2 METHODS

In this chapter, I describe the study settings, the data source, the data extracting method, the study variables, and a brief overview of the statistical methods that will be used in this thesis. The statistical analyses are described in more detail in individual analyses chapters. The data used in all analyses of this thesis were obtained from the South London and Maudsley NHS Foundation Trust, via the Clinical Record Interactive Search (CRIS) database.

2.1 SETTING

South London and Maudsley NHS Foundation Trust (SLAM) is the largest mental health trust in the UK.

The UK National Health Service (NHS) is organised into primary, secondary, and tertiary healthcare services (NICE, 2021). The primary care includes general practices, walk-in centres, and pharmacies. Serious or complex illnesses that cannot be effectively managed in primary healthcare are referred to secondary healthcare services. These are the more specialised hospital trusts such as SLAM, which is a mental healthcare trust that serves a catchment area. The tertiary healthcare services are even more specialised and typically covers a larger catchment area or even the whole country (NICE, 2021).

SLAM is a near-monopoly service provider of all aspects of secondary mental healthcare to over 1.3 million people of four London boroughs (Lambeth, Southwark, Lewisham, and

Croydon) (Stewart *et al.*, 2009; Perera *et al.*, 2016). SLAM's electronic health records give researchers the opportunity to design relatively comprehensive studies on people with disorders of interest within this defined geographic catchment area.

2.2 DATA SOURCE

The Clinical Record Interactive Search (CRIS) database is the de-identified version of SLAM's electronic health records (EHR). All the data used in this thesis were extracted from CRIS.

2.2.1 Clinical Record Interactive Search (CRIS) database

SLAM's Electronic Health Records (EHR) started in 2006. Two years later, the CRIS system was developed. CRIS provided a platform for the fully de-identified EHR of SLAM to become available to researchers to perform secondary data analysis within robust data security and governance framework (Stewart *et al.*, 2009). CRIS was approved as a de-identified patient data resource for secondary analysis by Oxfordshire Research Ethics Committee C (reference 18/SC/0372).

CRIS de-identifies patient records by masking the patients' names with ZZZZZZ and carers' names with QQQQQQ in all the text fields of the EHR. Figure 2.a shows the CRIS records where the patient's name is replaced with ZZZZZZ. In addition to masking names, patients' and carers' personal information is also truncated in CRIS, for example, the postcodes are truncated to the outer code only and the birth date information is truncated to only the month and year.

CRIS is a dynamic database of longitudinal information, which updates every 24 hours (Perera *et al.*, 2016). The dynamic nature of the CRIS database gives the researchers the

opportunity to examine dynamic cohorts, which is composed of comprehensive historic data together with the daily fresh update of new clinical information. As of January 2022, CRIS contains over 416,000 patient records extracted from over 38 million documents.

Typically, EHR is comprised of data in 2 formats: structured data and unstructured data. The CRIS database, which is the de-identified version of the EHR of SLAM, also includes these 2 formats. In addition, CRIS is enhanced with custom-built NLP algorithm results and linkage data. This makes the CRIS database a source of an unprecedented amount of information that is a valuable tool for mental health research.

In summary, the four types of data that are available via CRIS are structured data, unstructured data, custom-built NLP algorithm results and linkage data. Each of the data formats is described in detail below.

2.2.1.1 Structured data in CRIS

Structured data are characterised as those in which entry of data is constrained in some way, for example, information that is recorded via a drop-down menu. Examples of commonly used data that are retrieved from structured fields are gender, ethnicity, and date of birth. Structured fields are designed to standardise the information which is a valuable feature of data in research.

Figure 2.a shows an example of how structured data appears in CRIS. Table 2.b gives the list of all variables used in this thesis which were retrieved from structured fields of the CRIS database.

Unfortunately, the information in the structured fields is not often complete and much of the useful information are within the unstructured fields of the clinical notes.

2.2.1.2 Unstructured data in CRIS

Unstructured data refer to the information available in the free-text fields such as the clinical narratives written by physicians, nurses, and other healthcare providers. This includes clinical notes, correspondence, and inpatient events. The unstructured free-text fields of the CRIS database hold detailed records of almost every interaction the hospital has had with each patient, such as via face-to-face consultations, telephone calls, and email correspondence.

Even though the information in the unstructured free-text fields is more complete, extensive time-consuming manual reading or complex computational solutions are required to extract any useful information from it.

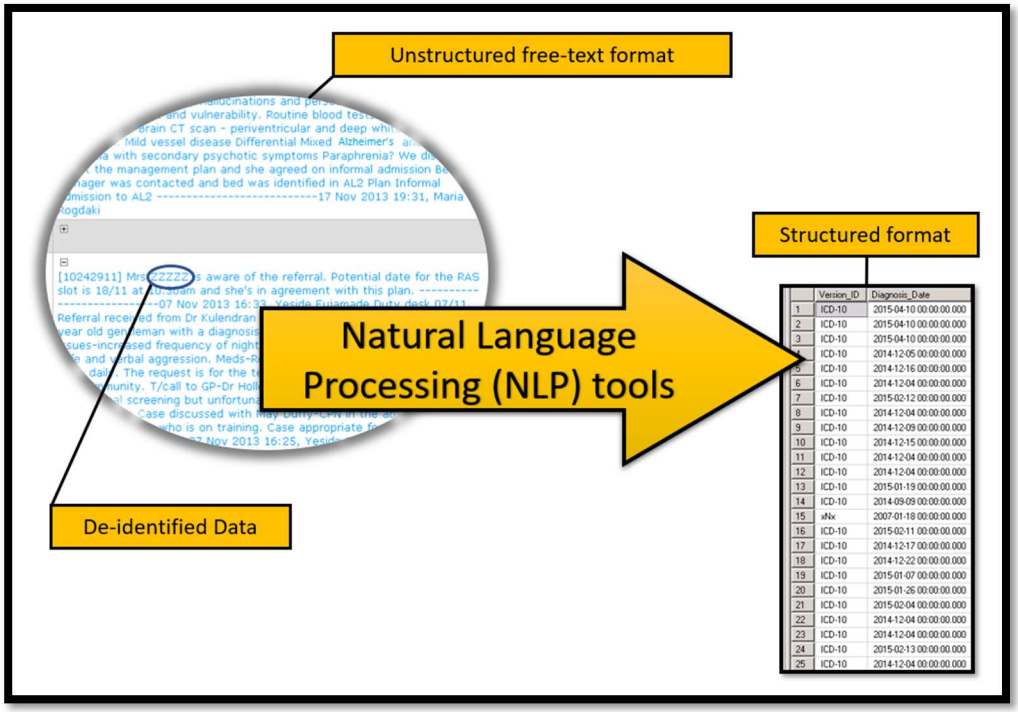
Figure 2.a shows an example of how unstructured data appears in CRIS. Table 2.b gives the list of all variables used in this thesis which were retrieved from unstructured fields of the CRIS database. In this thesis, all variables that needed to be extracted directly from unstructured data were extracted via the extensive time-consuming manual reading method. Computational solutions such as natural language processing (NLP) algorithms were not employed. However, the CRIS database includes pre-existing results from several NLP algorithms that were custom-built by a team of informaticians at CRIS. The pre-existing results of these algorithms were used in this thesis and are described below.

2.2.1.3 Custom-built NLP algorithm results in CRIS

Natural language processing (NLP) is a complex computational solution for extracting useful information from unstructured free-text documents and saving the results in the structured format (Chang *et al.*, 2011). This is illustrated in Figure 2.a. Researchers prefer to perform analysis on structured format data because it can be easily tabulated.

NLP algorithms outperform basic keyword searches style information extraction methods because the former also examines the linguistic aspects of the text, for example, it looks at the past tense and present tense features of a text and therefore can differentiate between the time-based characteristic of the following two phrases: 'is on clozapine' and 'previously took clozapine'. To help researchers exploit the strength of EHR data, the CRIS database is enhanced with pre-existing results of custom-built NLP algorithms.

Figure 2.a: Illustration of use of Natural Language Processing (NLP) and de-identification in CRIS data

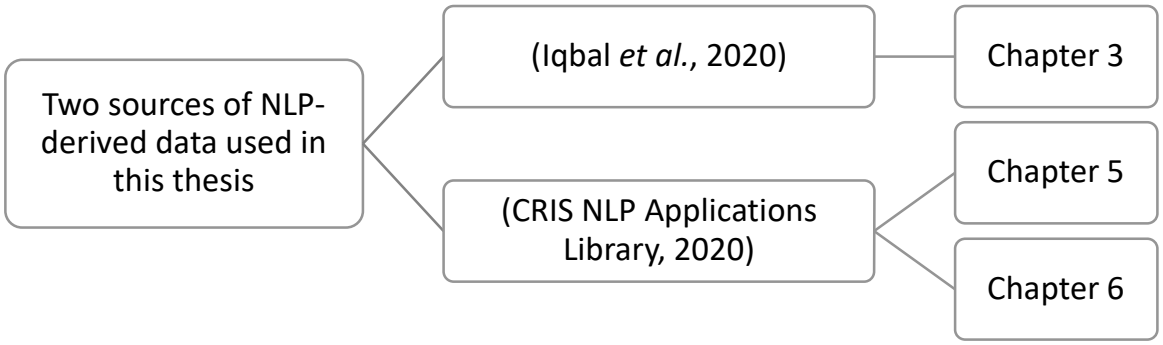


Natural Language Processing (NLP) is a computational technique that extracts information from unstructured free-text data and saves the results in a structured format. CRIS database, which is a de-identified resource also contains results from a custom-built NLP algorithm.

As illustrated in Figure 2.b, in this thesis, two sources of NLP-derived data were used: (i) algorithm developed by a former PhD student who used CRIS data to develop a clozapine-specific NLP algorithm (Iqbal *et al.*, 2020) (ii) NLP algorithm developed by a team of informaticians at CRIS (CRIS NLP Applications Library, 2020). Table 2.b gives the complete list of all variables used in this thesis that were the results of custom-built NLP algorithms. I was not involved in the development of these algorithms and used the results from the algorithm exactly as given.

As illustrated in Figure 2.b, all the NLP-derived data in chapter 3 was from (Iqbal *et al.*, 2020). No NLP-derived data were used in chapter 4. All the NLP-derived data used in chapters 5 and 6 were from (CRIS NLP Applications Library, 2020). Details of the NLP algorithm developed by Iqbal et al is given in chapter 3. Details of the NLP algorithms developed by the CRIS team is given below.

Figure 2.b: The two sources of NLP-derived data were used in this thesis



Two sources of NLP-derived data were used in this thesis: (i) algorithm developed by a former PhD student who used CRIS data to develop clozapine-specific NLP algorithm (Iqbal et al., 2020) (ii) NLP algorithm developed by a team of informaticians at CRIS (CRIS NLP Applications Library, 2020)

2.2.1.3.1 NLP algorithm developed by the CRIS team

There is a dedicated team of informaticians at CRIS who develop custom-built NLP algorithms using data from the unstructured fields within the CRIS database. These algorithms are developed, validated, and maintained by the CRIS team. The NLP-derived data are clearly labelled and stored alongside the structured and unstructured data so that researchers can access it seamlessly. All custom-built NLP algorithms in CRIS further improve the quality of its results by supplementing its NLP output with data from the structured fields, for example, the data from the ICD-10 diagnosis forms (structured format data) are used to supplement the diagnosis NLP algorithm (CRIS NLP Applications Library, 2020). Some examples of the custom-built NLP algorithm results present in CRIS are diagnosis, medication, BMI and smoking status of patients.

The NLP algorithms developed by the CRIS team are released with accompanying documentation that explains what each algorithm can and cannot do, and includes the relevant performance statistics of the algorithm (CRIS NLP Applications Library, 2020). To make sure that the results from the custom-built NLP algorithm are suitable for my research question, I studied these documentations carefully. The performance statistics that I used to make my decision were the precision and recall scores of each algorithm. The precision score is equivalent to the positive-predictive value while the recall score is equivalent to the sensitivity.

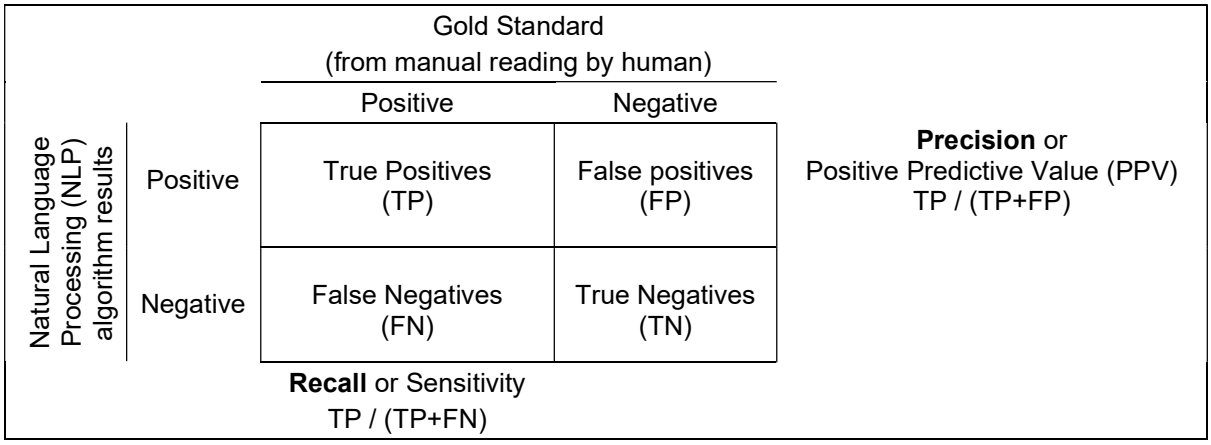
The precision and recall values are generated by comparing the results of the NLP algorithm against a gold standard (see Figure 2.c). In this case, the gold standard was created by a person in the CRIS team who manually read the same text and identified all the correct results that the NLP algorithm needed to identify in order to get the perfect score. Precision

and recall scores are proportions calculated by comparing the results of the algorithm against the expected results that are in the gold standard.

The precision score, which is equivalent to Positive Predictive Value (PPV) tells us that out of all results that were identified by the NLP algorithm, what proportion was correctly identified. The formula for precision is the number of true positive results identified by the NLP algorithm divided by the total number predicted as positive by the NLP algorithm, as illustrated in Figure 2.c.

The recall score, which is equivalent to sensitivity, tells us that out of all the results that were expected to be identified by the NLP algorithm, what proportion was identified. The gold standard plays the role of the benchmark for the expected results, as illustrated in Figure 2.c.

Figure 2.c: Framework for assessing the results of an NLP algorithm



An algorithm with a high precision score has a lower frequency of false positives and a high recall score have a lower frequency of false negatives. An algorithm with a 100% precision score and a 100% recall score has no false positive and no false negative. Such NLP

algorithms are ideal but almost never exist in reality. Depending on what is expected for each variable, a decision was made whether to compromise with having a high precision at the cost of missing some hits or having a high recall at the cost of including some false positives.

An example of the comprise between precision and recall is with the diagnosis NLP algorithm results. The results for the diagnosis algorithm were used as a key inclusion criteria for studies in chapter 5 and chapter 6. For this reason, I made sure that this algorithm had a high precision score, thus, minimising the frequency of false positives. The precision score for this algorithm is 100%, which is reassuring, meaning that there is an extremely low number or even possibly zero individuals in the cohort who were incorrectly diagnosed. The recall score for the diagnosis algorithm is 65%, which indicates that the data includes some false negatives. This means that there are individuals who should have been part of the cohort but were missed. Since this missing data was not restricted to either of the exposed or unexposed groups that I was comparing in my study, I accepted it as the missing data will be evenly distributed between the two outcome groups and thus not affect the results of my research. Therefore, the decision was made that compromise between this precision and recall score is acceptable, and this custom-built NLP algorithm is suitable for my research.

Results from custom-built NLP algorithms that were developed by the CRIS team were only used in chapters 5 and 6. Table 2.b gives the complete list of all variables used in these chapters that were the results of custom-built NLP algorithms. Data from four NLP algorithms were used: diagnosis, medication, smoking and body mass index (BMI) in both chapters. The precision and recall scores of the NLP algorithms are given in Table 2.a.

Table 2.a: Performance scores of custom-built algorithms that were used in this thesis

	Precision	Recall
Diagnosis	100%	65%
Antipsychotic Medication	88%	95%
Smoking		
Current smoker	79%	87%
Past smoker	68%	38%
Never smoked	72%	75%
Body mass index (BMI)	89%	78%

The performance statistics of the smoking algorithm was divided into three sub-categories: current smoker, past smoker and never smoked. The algorithm has quite a low recall score (38%) for the past smoker sub-category. This is because this information is not often clearly recorded in the clinical notes. We decided to use the results of this algorithm because all other scores for this algorithm were of an acceptable level.

If the precision or recall score of an NLP algorithm was too low to be acceptable, then I did not use it. For example, the recall score for the bronchitis NLP algorithm was 48%. A low recall score indicates a high proportion of false negatives. In this case, it means that there are a larger number of patients who have bronchitis but are labelled as otherwise by the algorithm. The bronchitis variable is a covariate in chapter 6. Using a covariate variable with a high proportion of false data can significantly affect the results of an analysis. For this reason, the results of this algorithm were not used in this thesis. The results of three other NLP algorithms were not used for the same reason: diabetes, asthma, and hypertension. These variables were extracted via the extensive time-consuming manual reading method (Table 2.b).

2.2.1.4 Clinical Data Linkage Service (CDLS) in CRIS

The Clinical Data Linkage Service (CDLS) is set up to enable external databases to be linked with the CRIS database in accordance with the research governance model required by the NHS for linking clinical data (Downs *et al.*, 2019). This governance model is from the 'Caldicott 2' report of the Department of Health Information Governance Review (Caldicott, 2013).

The main function of CDLS is to be a trusted third-party safe haven that ensures that the linking of confidential patient information is executed via a secure file transfer protocol and therefore guarantees the safety of the legal and ethical rights of patients and carers involved. Facilitated via CDLS, CRIS is linked to over 10 different local and national databases, for example, the local primary care database (Lambeth DataNet), the national cancer registry database (Thames Cancer Registry), the Hospital Episode Statistics (HES) and the Office for National Statistics (ONS) (Perera *et al.*, 2016).

In this thesis, three variables were retrieved from the CDLS data:

i) All-cause mortality information in chapter 6 was retrieved from the data linkage with the NHS spine.
ii) COVID-19 infection status in chapter 5 and 6 was retrieved from the data linkage with local hospitals (King's College Hospital and Princess Royal University Hospital)
iii) Clozapine blood monitoring data in chapter 4 was retrieved from the data linkage with Zaponex Treatment Access System (ZTAS) database

The variables are described in detail later in the chapter under section 2.4.

2.2.2 Zaponex Treatment Access System (ZTAS) database

The data from Zaponex Treatment Access System (ZTAS) is accessible through CRIS via data linkage. ZTAS is one of the UK's mandatory blood monitoring service providers. Clozapine patients treated at SLAM have their blood counts monitored by ZTAS (http://www.ztas.co.uk).

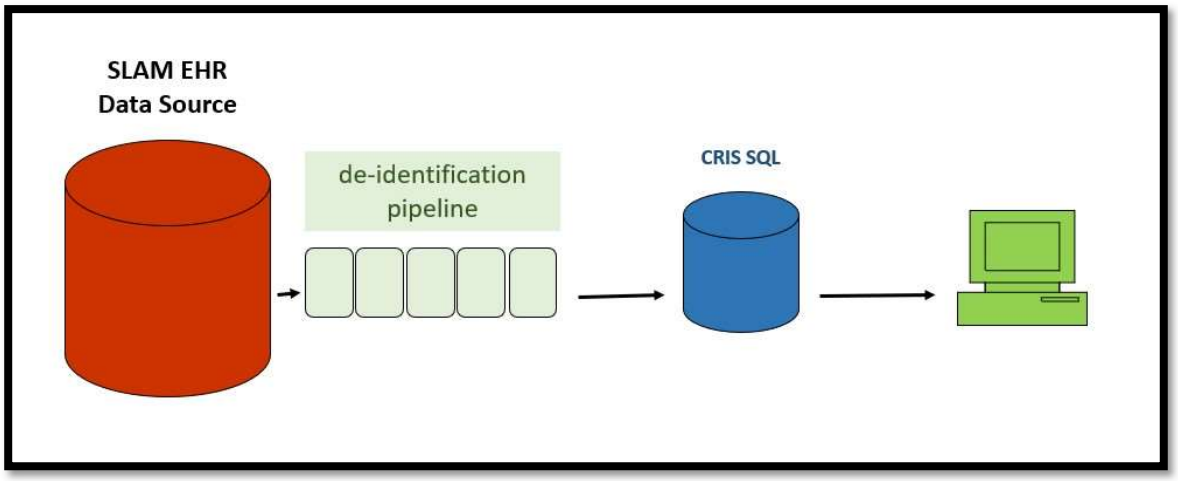
In the UK, the white blood count (WBC), absolute neutrophil count (ANC) and platelet count (PLT) are used for classifying the clozapine blood monitoring results into one of three risk categories: green, amber, red (Nielsen *et al.*, 2016). The ZTAS data accessed via CRIS contains almost 20 years of data, from May 2000 to October 2019. This gives CRIS users access to over 200,000 blood test results from over 2,000 SLAM patients. The data includes all WBC, ANC, PLT and category colours for each blood test.

The ZTAS blood test data was only used in chapter 4 and is described in detail there.

2.3 SQL – A DATA EXTRACTION TOOL

The CRIS data is stored in a SQL (Structured Query Language) database. As illustrated in Figure 2.d, the SLAM's electronic health records data are processed via the de-identification pipeline and the results are saved in an SQL database. This is the CRIS database. All data used in this thesis was extracted from the CRIS database using the SQL Server Management Studio version 15.0.

Figure 2.d: CRIS is a de-identified version of SLAM's electronic health records data. CRIS data is stored in an SQL database



The de-identification pipeline not only masks the names of the patients from the records but also assigns labels to all records corresponding to each patient with their unique identifier number, the BRC ID. The CRIS database is comprised of over 100 database tables and there is a BRC ID column in all of them. The BRC ID is unique per patient and is consistent across the database. The BRC ID is used to cross-reference and connect data from several tables is CRIS to create and extract a customised dataset for research.

As an example, I am providing in detail how data were extracted for chapter 4. This analysis compromised of data from three different database tables in CRIS: (i) clozapine blood test data table, (ii) SLAM pharmacy dispensary data table, (iii) SLAM clozapine clinic attendance data.

The clozapine blood test data was in a structured format. The source of this data was the Zaponex Treatment Access System (ZTAS) database that is accessible from CRIS (see section 2.2.2 for details). The SLAM pharmacy dispensary data was also in a structured format. The SLAM clozapine clinic attendance data was manually curated by reading the unstructured

free-text clinical notes, aided by the string search of the phrase similar to “attended clozapine clinic today”. Each row of data came labelled with the corresponding BRC ID.

The BRC ID was used to identify which data entry belongs to which patient. It also was the key to limiting the data to the patients of interest. The clozapine blood test data was used to set the inclusion criteria for the study cohort, meaning, the BRC IDs in this table was the basis of all patients that were to be included in the analysis. The other two tables were queried using the list of BRC IDs that existed in this table, thus their results were limited to these patients only. For this analysis, over 200,000 clozapine blood test data, over 300,000 pharmacy dispensary data and over 27,000 clinic attendance data entries were extracted. This data came from approximately 2,000 patients. SQL was used to extract the data and organise the data with respect to each patient. Details on how this was further processed are provided in chapter 4.

2.4 STUDY VARIABLES EXTRACTED FROM THE DATA SOURCE

Table 2.b shows the list of all variables used in this thesis alongside the information on the format the data was in before data extraction, meaning, whether was data was extracted from structured data, custom-built NLP algorithm or unstructured data.

Table 2.b: Variables used in each chapter. All variables were extracted from CRIS.

Variable Name	Source data format	Relevant chapter(s)			
		3	4	5	6
Clozapine-induced neutropenia	Structured data		✓		
	custom-built NLP algorithm	✓			
Clozapine treatment status	Structured data		✓		
	custom-built NLP algorithm	✓		✓	✓
COVID-19 infection	Structured data			✓	✓
COVID-related hospitalisation	Unstructured free-text data				✓
COVID-related intensive care treatment	Unstructured free-text data				✓
All-cause mortality	Structured data				✓
Age	Structured data	✓		✓	✓
Gender	Structured data	✓		✓	✓
Ethnicity	Structured data	✓		✓	✓
Welfare benefits	Structured data	✓			
Diagnosis	custom-built NLP algorithm			✓	✓
Smoker	custom-built NLP algorithm	✓		✓	✓
BMI or obesity	custom-built NLP algorithm			✓	✓
Inpatient	Structured data	✓		✓	
Contact with SLAM services	Structured data			✓	
Diabetes	Unstructured free-text data				✓
Asthma	Unstructured free-text data				✓
Bronchitis	Unstructured free-text data				✓
Hypertension	Unstructured free-text data				✓
Neighbourhood deprivation	Structured data				✓
Length of clozapine treatment	custom-built NLP algorithm	✓			

2.4.1 Different ways to extract the same variable

Table 2.b shows that two variables were extracted differently for different chapters of the thesis (clozapine-induced neutropenia and clozapine treatment status). This is a common feature of electronic health records data as there could be several ways to extract a variable.

2.4.1.1 Clozapine-induced neutropenia (structured data)

For chapter 4, clozapine-induced neutropenia, the outcome variable was extracted from structured data. The source of this data was the Zaponex Treatment Access System (ZTAS) database that is accessible from CRIS via the data linkage service. The ZTAS database and the data linkage service are described under section 2.2.2 and section 2.2.1.4, respectively.

ZTAS is SLAM's clozapine monitoring service. The data from ZTAS are the blood test results from the monitoring. The results are classified into one of three risk categories: green, amber, red (Nielsen *et al.*, 2016). The red results indicate the patient develops neutropenia.

Thus, for chapter 4, clozapine-induced neutropenia information was based on the red blood test results from the ZTAS data.

2.4.1.2 Clozapine-induced neutropenia (custom-built NLP algorithm)

For chapter 3, clozapine-induced neutropenia, the outcome variable, was extracted from the results of an NLP algorithm that was developed by (Iqbal *et al.*, 2020). This NLP algorithm text-mined and processed the mentions of neutropenia related keywords from unstructured free-text data in CRIS. I was not involved in the development of the algorithm and used the results from the algorithm exactly as given. The results were in a database

table with a new row for every clozapine-induced neutropenia event that was text-mined by the algorithm from the free-text data. Each row included a date for the event so that one can cross-check this with the clozapine treatment dates of the patient and verify that the neutropenia event occurred while was the patient was on clozapine treatment.

2.4.1.3 Clozapine treatment status (structured data)

For chapter 4, being on clozapine treatment was the inclusion criteria for the study. The start and stop dates of clozapine treatment were collated by combining information from three database tables of CRIS: (i) ZTAS blood monitoring data, (ii) SLAM pharmacy dispensary data table, (iii) SLAM clozapine clinic attendance data. The process of how the data were combined and the clozapine treatment dates was derived is described in chapter 4.

2.4.1.4 Clozapine treatment status (custom-built NLP algorithm)

For chapter 3, being on clozapine treatment was the inclusion criteria for the study. The start and stop dates of clozapine treatment were extracted from the results of an NLP algorithm that was developed by (Iqbal *et al.*, 2020). I was not involved in the development of the algorithm and used the results from the algorithm exactly as given. The results were in a database table with a new row for every clozapine treatment period information that was text-mined by the algorithm from the free-text data. The treatment period information included two dates: a treatment start date and a treatment end date.

For chapters 5 and 6, clozapine treatment status information was extracted from the results of the medication algorithm, a custom-built NLP algorithm that was developed within the

CRIS team (CRIS NLP Applications Library, 2020). For both chapters, clozapine treatment was the main exposure of interest and is described in detail below.

2.4.2 The main exposure of interest

Only chapters 5 and 6 have the main exposure of interest variable. In both chapters, people who were on clozapine treatment were designated as the exposed group. Those on any type or combination of antipsychotic treatment that did not include clozapine constituted the unexposed group. This information came from the results of the medication algorithm, a custom-built NLP algorithm. The precision and recall scores for the antipsychotics part of the medication algorithm are 88% and 90%, respectively (CRIS NLP Applications Library, 2020). Like all custom-built NLP algorithms in CRIS, this algorithm also further improves the quality of its results by combining its NLP results with data from the structured fields, such as the data from pharmacy prescriptions in the source record (CRIS NLP Applications Library, 2020).

The antipsychotics included in this thesis for the unexposed group were Olanzapine, Risperidone, Aripiprazole, Amisulpride, Paliperidone, Flupentixol, Haloperidol, Zuclopenthixol, Quetiapine, Fluphenazine, Piportil, Sulpiride, Lurasidone, Trifluoperazine, Chlorpromazine, Pipotiazine, Penfluridol, Droperidol, Pimozide, Thioridazine, Promazine, Ziprasidone Hydrochloride, Levomepromazine and Pericyazine.

2.4.3 Outcome Measures

All data for chapter 3 was from NLP-derived data from (Iqbal *et al.*, 2020), including the outcome measure. All other outcome measures in this thesis came from either structured data or manually curated by the reading of unstructured data.

For chapters 3 and 4, the outcome measure was clozapine-induced neutropenia. This variable was described previously under sections 2.4.1.2 (for chapter 3) and 2.4.1.1 (for chapter 4).

After the onset of the COVID-19 pandemic, I focused my PhD on investigating the effect of clozapine treatment on COVID-19 risk and severe outcomes. For chapter 5, the outcome measure was COVID-19 infection, and we tested whether clozapine treatment was associated with an increased risk of COVID-19 infection. Following this, we designed a follow-up study to investigate the associations between clozapine use and the severe outcome of COVID-19. For chapter 6, the outcome measure was COVID-related hospitalisation, COVID-related intensive care treatment and all-cause mortality.

2.4.3.1 COVID-19 infection

For chapter 5, the COVID-19 infection variable was the outcome variable. This data was extracted as structured format and was collated by combining information from the SLAM pathology lab results data, the presence of a clinician-entered alert on SLAM records: "COVID-19 positive" and information provided by local general hospitals (King's College Hospital and Princess Royal University Hospital) for COVID-19 related admissions.

For chapter 6, the COVID-19 infection data was used for the inclusion criteria of the study.

2.4.3.2 COVID-related hospitalisation

COVID-related hospitalisation was one of the three outcomes for chapter 6. This variable was extracted from unstructured data. The information was manually curated by reading the free-text clinical notes of each patient from the date of COVID-19 infection until a positive mention of hospitalisation or mention of recovery from COVID-19.

2.4.3.3 COVID-related intensive care treatment

COVID-related intensive care treatment was also an outcome for chapter 6. This variable was also extracted from unstructured data. The information was manually curated by reading the free-text clinical notes of each patient from the date of COVID-19 infection until a positive mention of intensive care treatment or mention of recovery from COVID-19.

2.4.3.4 All-cause mortality

All-cause mortality was also an outcome for chapter 6. This variable was extracted from structured data. In CRIS, the mortality information is populated on weekly basis via linkage with the NHS Spine. The NHS spine is a centralised database developed and maintained by the NHS for storing and sharing clinical data between the NHS trusts (Boyd A, Thomas R, 2018).

2.4.4 Covariates and explanatory variables

2.4.4.1 Age

Age was calculated from data that is routinely collected in structured format via the SLAM's Electronic Patient Records (EPR) Form. Patients are required to fill out the EPR form when they first enter SLAM's registry as a patient. In the CRIS database, this data is stored in a

table called ‘EPR Form’. For de-identification purposes, instead of the full date of birth information, CRIS only records the patient’s year and month of birth.

For this thesis, age was calculated from the year and month of birth. The age variable was used in chapters 3, 5 and 6. In chapter 3, this data was used for the age at the start of the first clozapine treatment. In chapter 5, this data was used for the age on the first day of the follow-up period of the study. In chapter 6, this data was used for the age at the time of COVID-19 infection.

2.4.4.2 Gender

The gender variable was also from the routinely collected structured data via the EPR Form. Gender variable was used in chapters 3, 5 and 6.

2.4.4.3 Ethnicity

The ethnicity variable was also from the routinely collected structured data via the EPR Form. In the EPR form, there are 14 sub-categories for ethnicity. In this thesis, this was collapsed into 4 sub-categories: White, Black, Asian and Others. The details are provided in chapters 3, 5 and 6.

2.4.4.4 Welfare benefits

The welfare benefits variable was also from the routinely collected structured data via the EPR Form. The purpose of this variable was to indicate the socioeconomic status of the patient. This variable was only used in chapter 3.

2.4.4.5 Diagnosis

The diagnosis variable came from the results of a custom-built NLP algorithm that was developed by the CRIS team. The precision and recall scores for the diagnosis algorithm are 100% and 65% respectively (CRIS NLP Applications Library, 2020). Like all custom-built NLP algorithms in CRIS, this algorithm also further improves the quality of its results by combining its NLP results with data from the structured fields, such as the data from ICD-10 diagnosis forms in the source record (CRIS NLP Applications Library, 2020). This combined data was the source for the diagnosis variable used in chapters 5 and 6.

In chapters 5 and 6, the diagnosis variable was utilized as the inclusion criteria to identify the individuals who were ever diagnosed with ICD-10 diagnoses of schizophrenia-spectrum disorders (F2*).

2.4.4.6 Smoker

The smoker variable came from the results of a custom-built NLP algorithm that was developed by the CRIS team. The smoker variable was used in chapters 5 and 6.

The NLP results of this variable segregated the data into three sub-categories: current smoker, past smoker and never smoked. In clinical records, a patient's smoking status can be recorded on multiple occasions. For this reason, for some patients, the smoking algorithm categorised some patients in more than one sub-group, meaning, for each time the smoking information is entered into the patient records, for some patients, there was an inconsistency in the information. For example, we had situations where the algorithm categorised a patient as a current smoker as well as never smoked. One possibility of this situation to arise is that the "never smoked" category was guessed by the healthcare

professional filling in a clinical form, without actually asking the patient. For our study, we needed each patient to only belong to one sub-category. When a patient belonged to more than one sub-category, we took the highest smoking status in the hierarchy “current smoker” > “past smoker” >“never smoked”. The precision (P) and recall (R) scores for each status of the smoking algorithm are as follows: for “current smoker” status, P=79% and R=87%; for “past smoker” status, P=68% and R=38%; for “never smoked” status, P=72% and R=75% (CRIS NLP Applications Library, 2020).

For chapter 5, we decided to accept the low recall score for “past-smoker” and all three sub-categories were included in the study. Chapter 6 was a follow-up study with a much smaller cohort size, so the smoking status was made a binary variable: current smoker vs non-current smoker.

2.4.4.7 BMI or obesity

The BMI and obesity variable came from the results of a custom-built NLP algorithm that was developed by the CRIS team. The overall precision and recall scores for the BMI algorithm are 89% and 78%, respectively (CRIS NLP Applications Library, 2020). To exclude erroneous values from the results of this algorithm, values were restricted to the range of 15 to 70 kg/m^2. From the algorithm results, the most recent BMI for each patient was used in chapters 5 and 6.

For chapter 5, the variable was called BMI. This was a categorical variable, and the BMI values were categorised into three subcategories: (1) underweight and healthy, (2) overweight, (3) obese. The BMI thresholds for the weight classification is provided in Table 2.c.

For chapter 6, the variable was called obesity. This was a much smaller cohort so there were only two subcategories: (1) obese (2) not obese.

Table 2.c: Body Mass Index (BMI) classifications (World Health Organization, 1995)

BMI, kg/m^2	Classification
< 18.5	Underweight
18.5 – 24.9	healthy
25-29.9	overweight
> 30	obese

2.4.4.8 Inpatient

The inpatient variable came from structured data that is routinely entered and updated by the healthcare providers at SLAM to keep track of patients receiving inpatient care. The inpatient variable was used in chapters 3 and 5, and in both studies, it was a binary variable. In chapter 3, this data was used to indicate if the patient was a SLAM inpatient at any time during the clozapine treatment. In chapter 5, this data was used to indicate patients' SLAM inpatient status on the first day of the follow-up period of the study.

2.4.4.9 Contact with SLAM services

The contact with SLAM services variables was manually curated by combining information from several structured data fields, such as the presence of information that indicated that the patient attended face-to-face consultations, phone calls, email correspondences or any other form of inpatient or outpatient communication. This variable was only used in chapter 5 and details are provided there.

2.4.4.10 Diabetes

The diabetes variable was extracted from unstructured data. The information was manually curated by reading the free-text clinical notes of each patient. Since there was not enough time to read all the clinical notes of all the patients in the cohort, the aid of search strings was used. We retrieved all unstructured free-text fields of all the patients in the cohort that included the search keywords diabetes and diabetic. Over 1000 unstructured free-text fields were retrieved and manually read to curate the values for the diabetes variable. This variable was only used in chapter 6.

2.4.4.11 Asthma

The asthma variable, which was also only used in chapter 6 was manually curated from unstructured data using the same methods described for extracting the diabetes variable under section 2.4.4.10.

The searched keywords were asthma and asthmatic.

2.4.4.12 Bronchitis

The bronchitis variable, which was also only used in chapter 6 was manually curated from unstructured data using the same methods described for extracting the diabetes variable under section 2.4.4.10.

The searched keywords were bronchitis, COPD, chronic obstructive pulmonary disease (which is what COPD stands for) and chronic obstructive airway disease.

2.4.4.13 Hypertension

The hypertension variable, which was also only used in chapter 6 was manually curated from unstructured data using the same methods described for extracting the diabetes variable under section 2.4.4.10.

The searched keywords were hypertension and high blood pressure.

2.4.4.14 Neighbourhood deprivation

The neighbourhood deprivation variable was extracted from structured fields in CRIS. This was only used in chapter 6 and details are provided there.

2.4.4.15 Length of clozapine treatment

The length of the clozapine treatment variable came from the results of a custom-built NLP algorithm that was developed by (Iqbal *et al.*, 2020). This variable was only used in chapter 3 and details are provided there.

2.5 STATISTICAL ANALYSIS

The statistical analysis for chapter 3 was performed using the glm package in R (version 3.5.3). The statistical analysis for chapters 4, 5 and 6 was performed using STATA for Windows version 15.1. Full details of the statistical analysis methods are described in each chapter and an overview is provided here.

In chapter 3, we used logistic regression to determine the association between explanatory variables and clozapine-induced neutropenia. Logistic regression was used because the outcome was a binary variable.

In chapter 4, we used the Kaplan-Meier survival curve to determine the time to clozapine-induced neutropenia. Time zero was the beginning of the clozapine treatment period.

In chapter 5, we used Cox proportional hazard models to calculate hazard ratios for COVID-19 positive status, in clozapine treated patients versus those treated with other antipsychotics.

In chapter 6, we used logistic regression to test for association between clozapine treatment and severe outcomes of COVID-19, namely, hospitalisation, intensive care treatment and mortality). Odds ratios were calculated by comparing clozapine treated patients to those treated with other antipsychotics. Logistic regression was used instead of Cox proportional hazard models because we did not have a precise date for time zero, which was the date of COVID-19 infection. We had the date for the date of swab or date of the test result of a COVID-19 test, which was not sufficient to identify the true date of the infection.

CHAPTER 3

3 CHALLENGES OF USING ELECTRONIC HEALTH RECORDS DATA FOR RESEARCH

3.1 ABSTRACT

Background

Clozapine is an antipsychotic medicine that is used to treat patients with schizophrenia after other treatments have failed. Unfortunately, clozapine is associated with a rare but potentially fatal adverse drug reaction (ADR) called neutropenia, a type of blood dyscrasia. There are currently no clinical predictors for clozapine-induced neutropenia.

South London and Maudsley NHS Foundation Trust (SLAM) is the largest mental health trust in the UK. SLAM's Electronic Health Records (EHR) is a powerful resource for performing research studies related to psychiatric disorders. Iqbal and colleagues recently used SLAM's EHR data to develop the Clozapine Adverse Drug Reaction (CLZ-ADR) algorithm. The CLZ-ADR algorithm identifies clozapine treatment dates and clozapine ADR dates in clozapine-treated SLAM patients and stores the results in database tables.

Aims

To use the dataset produced by the CLZ-ADR algorithm to build a model for identifying predictors of clozapine-induced neutropenia in SLAM patients.

Methods

The study combined information from 5 different database tables from SLAM's EHR. Exploratory data analysis was performed. Logistic regression models were constructed to

identify the explanatory variables that were independently associated with clozapine-induced neutropenia.

Results

The cohort of the study was comprised of the 1760 patients who were identified to have received clozapine treatment in SLAM between 2007 and 2017.

To build the model, seven explanatory variables were selected, ethnicity, gender, welfare benefits status, smoking status, inpatient status, age at start of first clozapine treatment and length of clozapine treatment in days. Logistic regression model results showed that 3 variables were associated with clozapine-induced neutropenia: age at first clozapine treatment (OR=0.98, 95% CI 0.97 - 0.99), black ethnicity (OR=2.48, 95% CI 1.89-3.27) and smokers (OR=1.49, 95% CI 1.13-1.96). We also found that the CLZ-ADR algorithm only reports clozapine treatment that was 90 days or longer. This filtering threshold was pre-set by the algorithm to reduce false positives. Since most the clozapine-induced neutropenia cases are expected to occur in the first few months of clozapine treatment, this pre-set threshold set by the CLZ-ADR algorithm became a major barrier for using its data to build a predictive model for clozapine-induced neutropenia.

Conclusion

In conclusion, there are some challenges in working with EHR data which impacted the ability of this study to achieve its aim of identifying predictors of clozapine-induced neutropenia in SLAM patients. One minor oversight during study design can lead to irrecoverable limitations in the analysis. That said, with a cautious study design and thorough awareness of the challenges of working with EHR data, EHR is an extremely useful resource for research.

3.2 INTRODUCTION

Electronic Health Records (EHR) are a rich real-world dataset resource containing clinical data that can provide a valuable platform for performing retrospective and observational studies. EHR include structured data as well as free-text clinical notes. Structured data refers to data organised in a table-style format and can be used for research. The free-text clinical notes refer to the clinical information written in words, in a narrative format especially by nurses and clinicians to keep detailed records of their consultations with patients. The lack of structure in the free-text clinician notes is a major obstacle to fully utilising EHR for research. Even though the free-text clinical notes store the most robust and potentially complete clinician information about a patient, this information is embedded deep within the narratives and is difficult to extract without reading manually. Several studies have shown that important clinical information gets lost when free-text clinical notes are not used (Tate *et al.*, 2009, 2011; Joling *et al.*, 2011; Ford *et al.*, 2013; Capurro *et al.*, 2014; Morrison, 2020).

Natural Language Processing (NLP) techniques are a text-mining approach to turn free-text narratives into structured data that can be used in research studies (Jackson *et al.*, 2017). NLP algorithms are more elaborate than basic keyword searches because they evaluate the linguistic aspect of a phrase, for example, temporal modifiers (e.g., “is currently on clozapine” versus “was previously on clozapine”). There is an ongoing effort to develop NLP solutions to extract information from free-text clinical notes in several clinical settings. One such setting is South London and Maudsley NHS Foundation Trust (SLAM).

SLAM, the largest mental health trust in the UK, moved from a paper-based records system to EHR in 2006. Two years later, SLAM developed a Clinical Record Interactive Search (CRIS) system to create a platform with a fully deidentified case register so researchers can perform secondary analysis within a robust data security and governance framework (Stewart *et al.*, 2009). CRIS not only contains specific structured fields but also results from custom-built NLP algorithms built by a designated team of NLP specialists. In addition, CRIS contains results from other text-mining approaches built by researchers who use CRIS data. One of such text-mined results is from the Clozapine Adverse Drug Reaction (CLZ-ADR) algorithm (Iqbal *et al.*, 2020).

I co-authored the CLZ-ADR algorithm paper, which was published in December 2020 (Iqbal *et al.*, 2020). An adverse drug reaction (ADR) is an unwanted or harmful response to a drug. The CLZ-ADR algorithm is a text mining approach for identifying clozapine treatment dates and clozapine ADR dates from free-text clinical records of CRIS. I performed the manual validations of the text-mined results. The CLZ-ADR algorithm is the cornerstone of this study as all text-mined data used in this chapter are extracted from the results generated by it.

The aim of this study is to (1) develop a prediction model for adverse drug reactions on clozapine, starting with data generated by the CLZ-ADR algorithm (2) improve the model by adding other sources of data available in CRIS.

3.2.1 Clozapine-induced neutropenia

Clozapine is an antipsychotic medicine that is used to treat patients with schizophrenia after other treatments have failed. Even though 23% of schizophrenia patients do not respond to any antipsychotic medications other than clozapine, clozapine is an underutilized

medication (Siskind *et al.*, 2021). This is because clozapine is associated with a rare but potentially fatal ADR called neutropenia. There are currently no clinical predictors for this clozapine-induced neutropenia.

Neutropenia is a blood dyscrasia characterised by a severe reduction in neutrophils in the blood. Neutrophils are a type of white blood cells and severe reductions in neutrophils increase the risk of fatal infections. The normal range for neutrophil count is between $2x10^9$/L and $7.5x10^9$/L. Neutropenia occurs when the neutrophil count drops to less than $1.5x10^9$/L. Agranulocytosis, also knowns as severe neutropenia, occurs when the neutrophil count drops to less than $0.5x10^9$/L. In clozapine-treated patients, the prevalence of neutropenia and agranulocytosis is reported to be 3.8% and 0.4%, respectively (Myles *et al.*, 2018; Li *et al.*, 2020). Unfortunately, there are no known clinical predictors for clozapine-induced neutropenia or agranulocytosis. In this study, I aim to use the data generated by the CLZ-ADR algorithm to build a predictive model for clozapine-induced neutropenia.

3.3 METHOD

3.3.1 Data Source and Ethics Statement

CRIS is a deidentified version of SLAM's EHR (Stewart *et al.*, 2009). SLAM caters to all secondary mental health care needs of over 1.3 million people of four London boroughs (Lambeth, Southwark, Lewisham, and Croydon). CRIS was approved for use as a deidentified data resource for secondary analysis by Oxfordshire Research Ethics Committee C (reference 18/SC/0372).

CRIS is comprised of hundreds of tables, thousands of columns, and millions of data points. The information in different tables can be combined based on the unique patient identifier numbers that is mentioned alongside each data point. CRIS includes structured data, free-text clinical narratives and results from text-mining algorithms. The results from the text-mining algorithms are extracted information from the free-text clinical narratives in structured formats that can be used in research studies. One such text-mining algorithm is the CLZ-ADR algorithm (Iqbal *et al.*, 2020).

Data from the CLZ-ADR algorithm was used in this study. The CLZ-ADR algorithm extracts several types of information from CRIS and organises the results into several database tables. Four of the CLZ-ADR algorithm results database tables are used in this study.

This study also included exploring data available in CRIS on the patient's demographic information. Database table 5 stores unprocessed data on patient demographics in a structured format.

In total, this study combined information from 5 different database tables. All 5 tables only contain structured format data. **Table *3.a*** shows the features of the 5 database tables used in this study - the cohort defining data, outcome measure data, smoking data, inpatient data and demographic data.

Table 3.a: This study combines information from five different database tables

	Database Table Name	Source	Type
1	Cohort defining data	Results of the CLZ-ADR algorithm (Iqbal *et al.*, 2020).	Text-mined
2	Outcome Measure	Results of the CLZ-ADR algorithm (Iqbal *et al.*, 2020).	Text-mined
3	Smoking data	Results of the CLZ-ADR algorithm (Iqbal *et al.*, 2020).	Text-mined
4	Inpatient data	Results of the CLZ-ADR algorithm (Iqbal *et al.*, 2020).	Text-mined
5	Demographic data	Unprocessed data from patient records table in CRIS	Structured data in CRIS

3.3.1.1 Database Table 1 – Cohort defining data

The cohort of the study was all patients receiving clozapine treatment in SLAM between 2007 and 2017. This information was extracted from free-text clinical notes using text-mining approaches via the CLZ-ADR algorithm (Iqbal *et al.*, 2020).

Since the CLZ-ADR algorithm text-mines several types of information, the results are organised into several database tables. The database table that stored the clozapine treatment dates information also included information on gender, ethnicity, age at the start of the first clozapine treatment and length of clozapine treatment. These variables were included in the analysis.

The gender data was provided as a categorical variable with 2 subcategories: Male, Female

The ethnicity data was provided as a categorical variable with 4 subcategories: White, Black, Asian and Other.

The age at start of first clozapine treatment was provided as a numeric variable. It was calculated from the patient's date of birth and the first date of clozapine treatment identified by the CLZ-ADR algorithm.

The length of clozapine treatment was also provided as a numeric variable. This is the total number of days the patient was on clozapine treatment.

3.3.1.2 Database Table 2 – Outcome Measure data

The outcome of interest was clozapine-induced neutropenia. This included data on severe neutropenia, which is also known as agranulocytosis. The outcome measure was also identified by text-mining approaches via the CLZ-ADR algorithm (Iqbal *et al.*, 2020).

Clozapine-induced neutropenia is an adverse drug reaction (ADR). The ADR identification part of the CLZ-ADR algorithm is the ADEPt algorithm (Iqbal *et al.*, 2017). The CLZ-ADR algorithm and the ADEPt algorithm were both developed in the CRIS environment by Ehtesham Iqbal (Iqbal *et al.*, 2017, 2020).

The database table that stored ADR information was in the format of one row per each ADR incident, meaning there were as many rows per person as the number of times each person experienced an ADR. Each row of the database table included information on the date of the ADR and the start and ends dates of the corresponding clozapine treatment. There was information on several different clozapine-induced ADRs, including neutropenia and severe neutropenia (agranulocytosis).

I generated a binary variable using the clozapine-induced neutropenia or severe neutropenia data. Patients who had experienced clozapine-induced neutropenia or severe neutropenia during their treatment with clozapine were categorised into the category "1" and the rest were categorised as "0". This binary variable became the outcome variable of the predictive model and was named clozapine-induced neutropenia.

3.3.1.3 Database Table 3 – Smoking data

The smoking database table contained information on whether the patient smoked during clozapine treatment or not. This data was generated by text-mining approaches via the CLZ-ADR algorithm (Iqbal *et al.*, 2020).

3.3.1.4 Database Table 4 – Inpatient data

The inpatient database table contained information on whether the patient was an inpatient during clozapine treatment or not. This data was generated by text-mining approaches via the CLZ-ADR algorithm (Iqbal *et al.*, 2020). This database table only contains data on inpatients and therefore any patients who were not in this dataset were classified as "not inpatient".

3.3.1.5 Database Table 5 – Demographics data

The demographic database table contains information that was collected via the Electronic Patient Record (EPR) Form. The EPR form is used during hospital admission to collect data in pre-defined values form, therefore the patient information collected via the EPR form is saved in a structured format in CRIS. According to many CRIS database researchers, this table is the first table researchers explore for patient demographics data as the data is available in a structured format.

3.3.2 Selecting explanatory variables from Database Table 5 – Demographics data

In addition to the results from the CLZ-ADR algorithm, I wanted to include patients' demographic-related variables in the predictive model. Database table 5 (demographic data) stores unprocessed patient demographic data collected directly from patients via a form. All variables in database table 5 are in a structured format. Each variable in database table 5 was explored in detail so that relevant demographic-related variables can be selected for the predictive model. Summaries of each variables is shown in **Table *3.b***.

Database table 5 contains 19 types of demographic data on all SLAM patients, including date of birth, gender, ethnicity, disabilities, housing situation, employment, living status in the UK, country of origin, religion, and language.

In addition to including missing data, the table also included multiple ways to indicate missing data, for example, “NULL”, “Not Known”, “Not known”, “Unknown”, “Not Applicable”, “Other”, “Not Disclosed”. The fields in the tables with these values were recoded to missing. The number of missing data points in each variable was calculated and the variables with more than 25% of missing data were removed.

The categorical variables were tabulated and the percentage of data in each category was calculated. The proportions of data in each category were assessed and variables with more than 80% of data in one category were removed.

Variables such as gender and ethnicity also exist in database table 1 (cohort defining data). The corresponding duplicated variables were identified and quality checked for consistency. The variables in the database table 5 (demographic data) that were deemed duplicates were removed.

It is typical of variables sourced from EHR to have a high proportion of missing data or have a high proportion of data in one category or contain duplicate variables. These typical characteristics of EHR data were thoroughly checked in the data from database table 5 (demographic data) as this was the only dataset in the study that did not come from the CLZ-ADR algorithm and therefore was not previously quality-checked.

In summary, the variables that belonged to any of the following three groups were removed and not used in the analysis:

(1) Variables that had >25% missing data

(2) Categorical variables that had >80% of data is one sub-category

(3) Variables that were duplicate or redundant variables.

The details of the variables from database table 5 (demographic data) that were removed from the analysis are shown in **Figure *3.a***.

3.3.3 Combining data from 5 tables

For the statistical analysis, I needed to tabulate the data in the format of 1 row per patient. The columns of the tables will correspond to the variables of the predictive model. Of the 5 database tables used in this study, all tables were in 1 row per patient format, except database table 2 (outcome measure data).

The database table 2 (outcome measure data) was used to generate the outcome measure, clozapine-induced neutropenia. The details on how this binary variable was generated were under section 3.3.1.2.

The database table 5 (demographic data) contains data on all patients who received treatment in SLAM since the records started in CRIS. The list of patient ids from database table 1 (cohort defining data) were used to generate a subset of database table 5 to limit it to only the patients in this study.

3.3.4 Statistical analysis

Statistical analysis was performed to test how the explanatory variables predicted clozapine-induced neutropenia.

The explanatory variables included the 6 explanatory variables that came from the results of the CLZ-ADR algorithm: age at start of first clozapine treatment, gender, ethnicity, smoking status, inpatient status, and length of clozapine treatment in days. In addition, the explanatory variables included variables from database table 5 (demographic data) that passed the selection criteria.

The outcome variable came from the results of the CLZ-ADR algorithm and was called clozapine-induced neutropenia. This data was the binary categorical format of the patients who experienced clozapine-induced neutropenia and patients who did not experience clozapine-induced neutropenia, and so logistic regression was used for modelling.

3.3.4.1 Exploratory Data Analysis

Exploratory data analysis was performed on each explanatory variable that was used in logistic regression modelling.

The continuous variables were examined using histograms and Mann-Whitney U test to explore the difference in the distribution of the variables between patients who experienced clozapine-induced neutropenia and patients who did not experience clozapine-induced neutropenia (Figure 3.b and *Table 3.c*)

The categorical variables were examined using bar charts and Fisher's exact test to explore the difference in the distribution of the variables between patients who experienced

clozapine-induced neutropenia and patients who did not experience clozapine-induced neutropenia (Figure 3.c and **Table 3.c**).

3.3.4.2 Logistic Regression

Logistic regression models were constructed on the explanatory variables defined above to identify variables independently associated with clozapine-induced neutropenia.

All analysis was performed with the glm package in R (version 3.5.3).

Table 3.b: Summaries of data in database table 5 - the unprocessed patient demographic information stored in a structured format in CRIS.

Dataset 5 –Demographics Data (n=1760)	
Gender	Males: 1167 Females: 593
Date of Birth	*[list of Dates]*
Ethnicity	British:697 Any other black background:266 Caribbean:223 African:214 White Irish: 30 Any other white background: 93 Mixed Race – W & B Caribbean: 32 Mixed Race – W & B African: 12 British Indian: 23 British Bangladeshi: 10 British Chinese: 7 (other categories):147 NA: 6
Welfare Benefits	Yes: 614 No :1146
Mobility Problem	Yes: 45 No:1714 No Known Problems: 1
Interpreter Needed	Yes: 37 No :1723
Overseas Visitor	Yes: 17 No :1743
Visual Impairment	Registered Partially Blind: 2 Unregistered Blind: 1 Unregistered Partially Blind: 5 No Known Problems: 1 None :1727 NA: 24
Hearing Impairment	No Known Problems: 1 None:1722 Registered Deaf: 4 Unregistered Hard of Hearing: 5 NA: 28
Marital Status	Single:1467 Married: 94 Divorced: 61 Separated: 42 Married/Civil Partner: 32 (other categories): 26 NA: 38
Has A Twin	No: 400 Yes, Alive: 12 NA:1348
Asylum Seeker	No: 273 Indefinite Leave to Remain: 16 Leave to Remain: 9 Application Pending: 7 Illegal Immigrant: 3 (other categories): 4 NA:1448
Employment	Unemployed:876 Full Time Student: 33 Full Time Student - School age: 20 Retired: 16 Volunteer: 13 Part Time employment: 13 Paid Employment: 13 (other categories): 6 NA:770
First Language	English: 1197 Somali: 9 Arabic: 5 Tamil: 5 French: 4 (other categories): 38 NA: 502
Country Of Origin	United Kingdom:530 England: 35 Nigeria: 31 Jamaica: 30 Somalia: 16 (other categories):173 NA:945
Lives With	Alone: 360 Parents: 87 Mother: 59 Partner: 53 Other Relatives: 36 (other categories): 43 NA:1122
Housing Status	Council Tenant: 277 Homeless: 36 Nursing/Residential: 64 Owner: 39 Private Tenant: 113 Trust: 26 NA:1205
Religion	Church of England: 92 Other Christian: 78 Christian: 68 Roman Catholic: 74 Other Protestant: 64 Muslim: 59 Atheist/Agnostic: 33 Hindu: 14 (other categories): 51 NA:1227
Residence	Lambeth: 156 Southwark: 70 Lewisham: 65 Croydon: 60 Surrey: 7 (other categories): 13 NA:1389

Note that some categorical variables had a high number of subcategories and the number of patients that have data in a category that is not shown in the table is shown as (other categories). The missing values in each variable are shown as NA, which stands for "Not Available".

3.4 RESULTS

From the results of the CLZ-ADR algorithm, 1760 patients were identified to have received clozapine treatment in SLAM between 2007 and 2017. This was the cohort of the study. The explanatory variables came from 4 different database tables. The outcome variable was the clozapine-induced neutropenia which came from a single database table.

Gender information exists in database table 1 and database table 5. Quality checks showed that the 2 sources of data were identical therefore one of the variables was dropped.

Ethnicity was another variable that existed in database table 1 and database table 5. Unlike the gender variable, the two sources of ethnicity data were not identical, but quality checks showed that there was consistency between the two. The ethnicity data in database table 5 had 17 categories. Database table 1 had the same data collapsed into 4 categories so this collapsed version was used in this study.

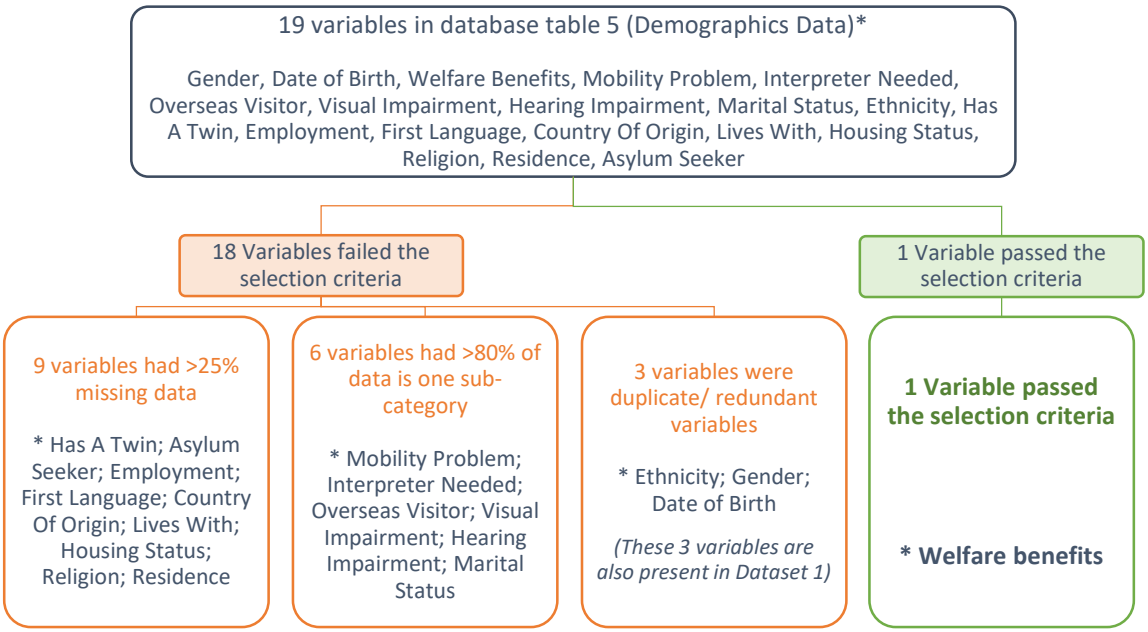
The data of birth information in database table 5 (demographics data) was truncated for de-identifying purposes. This analysis was performed via CRIS data, which is a deidentified version of SLAM's EHR. In order to keep the patients' identity hidden, the date of birth of all patients is modified to the 1st day of their date of birth month, for example, 10/03/1984 is modified to 01/03/1984.

The lowest value for the length of the clozapine treatment variable is 90 days because that was the pre-defined lower threshold set by the CLZ-ADR algorithm. The maximum value for this variable is 3803 days, which is over 10 years. This shows this dataset includes chronic patients who have remained on clozapine treatment for a substantial length of time.

Table *3.b* shows the summaries of the data in database table 5 (demographics data). Database table 5 was the only data used in the study that did not come from the CLZ-ADR algorithm. Since this table stores unprocessed patient demographic data collected directly from patients via a form, the variables of database table 5 were thoroughly explored. There were 19 variables in database table 5 (demographics data) and the summaries of this data is shown in **Table *3.b***.

Figure 3.a shows that 18 out of the 19 variables in database table 5 were removed from the analysis. The following 9 variables were removed because they had more than 25% missing data: Has A Twin; Asylum Seeker; Employment; First Language; Country Of Origin; Lives With; Housing Status; Religion; Residence. The following 6 categorical variables were removed because they had more than 80% data in one category: Mobility Problem; Interpreter Needed; Overseas Visitor; Visual Impairment; Hearing Impairment; Marital Status. The following 3 variables were removed because they were deemed redundant as they were duplicates of data present in database table 1: Ethnicity; Gender; Date of Birth. The one variable from database table 5 (demographics data) that was included in the analysis was welfare benefits.

Figure 3.a: Selecting explanatory variables from database table 5 based of typical EHR characteristics



** Some typical characteristics of EHR data are present in the variables of Database Table 5. These characteristics were assessed and variables that did not pass the selection criteria were removed*

There were 7 explanatory variables that were selected to build the model were: age at start of first clozapine treatment, gender, ethnicity, smoking status, inpatient status, welfare benefits status and length of clozapine treatment in days. Apart from welfare benefits information, all data came from the results of the CLZ-ADR algorithm. Welfare benefits information was included as the demographic data that indicated the socioeconomic status of the patient. The inpatient status was included to indicate the severity of the illness of the patient.

Table 3.c shows the summary statistics of the data used in the modelling. The explanatory variables comprised 5 categorical variables and 2 continuous variables. The 2 continuous variables were age at start of first clozapine treatment and length of clozapine Treatment in days. The 5 categorical variables were ethnicity, gender, welfare benefits status, smoking

status and inpatient status. Apart from ethnicity, all categorical variables were binary variables and indicated as 1 if present and 0 if absent (for gender 1 was female and 0 male).

The final dataset compromised of no missing data as all patients had information on all the variables. Exploratory data analysis was performed on each explanatory variable that was used in logistic regression modelling.

Histograms in **Figure *3.b*** show the distribution of patients who experienced clozapine-induced neutropenia and patients who did not experience clozapine-induced neutropenia within the continuous explanatory variables used for logistic regression modelling. For the age at start of first clozapine treatment, the distribution was modestly right-skewed for patients who experienced clozapine-induced neutropenia, indicating that patients who started clozapine treatment at a younger age may be more likely to experience clozapine-induced neutropenia. For the length of treatment, the distributions were seen to be similar between the two groups of patients.

Bar charts in Figure 3.c show the proportions of patients who experienced clozapine-induced neutropenia and patients who did not experience clozapine-induced neutropenia within each category of the categorical explanatory variables used for logistic regression modelling. There was a modestly higher incidence of clozapine-induced neutropenia in patients who smoked during clozapine treatment, were inpatients during clozapine treatment or had black ethnicity. The incidence of clozapine-induced neutropenia was similar in males and females. The incidence of clozapine-induced neutropenia was similar in patients who were and were not on welfare benefits.

Mann-Whitney U test results in *Table 3.d* showed that when comparing in patients who experienced clozapine-induced neutropenia with and those who did not, 1 continuous explanatory variable, the age at first clozapine treatment was lower in patients with

clozapine-induced neutropenia (p=3.7e-06). Patients who experienced clozapine-induced neutropenia had a lower mean for the age at first clozapine treatment (36.5 ± 11.9) compared to those who did not (40.0 ± 12.0).

Fisher exact test results in *Table 3.d* showed that when comparing patients who experienced clozapine-induced neutropenia and patients who did not experience clozapine-induced neutropenia, 3 categorical explanatory variables had significant p-values, smokers (p=1.4e-04), ethnicity (p=1.5e-04) and inpatient (p=8.2e-04). A higher proportion of patients who experienced clozapine-induced neutropenia smoked during their clozapine (68.3%), compared to those who did not (56.9%). A higher proportion of patients who experienced clozapine-induced neutropenia were of black ethnicity (52.6%), compared to those who did not (36.2%). A lower proportion of patients who experienced clozapine-induced neutropenia were of white ethnicity (33.5%), compared to those who did not (50.5%). A higher proportion of patients who experienced clozapine-induced neutropenia were inpatients (43.0%%), compared to those who did not (33.1%).

Logistic regression model results in **Table 3.e** showed that 3 variables were associated with clozapine-induced neutropenia: age at first clozapine treatment (p=6.98e-04), black ethnicity (p=8.72e-11) and smokers (p=4.52e-03).

The odds ratio for age at first clozapine treatment (OR=0.98, 95% CI 0.97 - 0.99) indicates that lower age at first clozapine treatment is associated with an increase in the probability of clozapine-induced neutropenia. This OR is so close to 1 because the OR represents the change in risk for each increasing year of age. Across a 10-year period, the risk of clozapine-

induced neutropenia will be substantially reduced with an OR of 0.83 (95% CI 0.74 -0.93) compared to starting 10 years earlier.

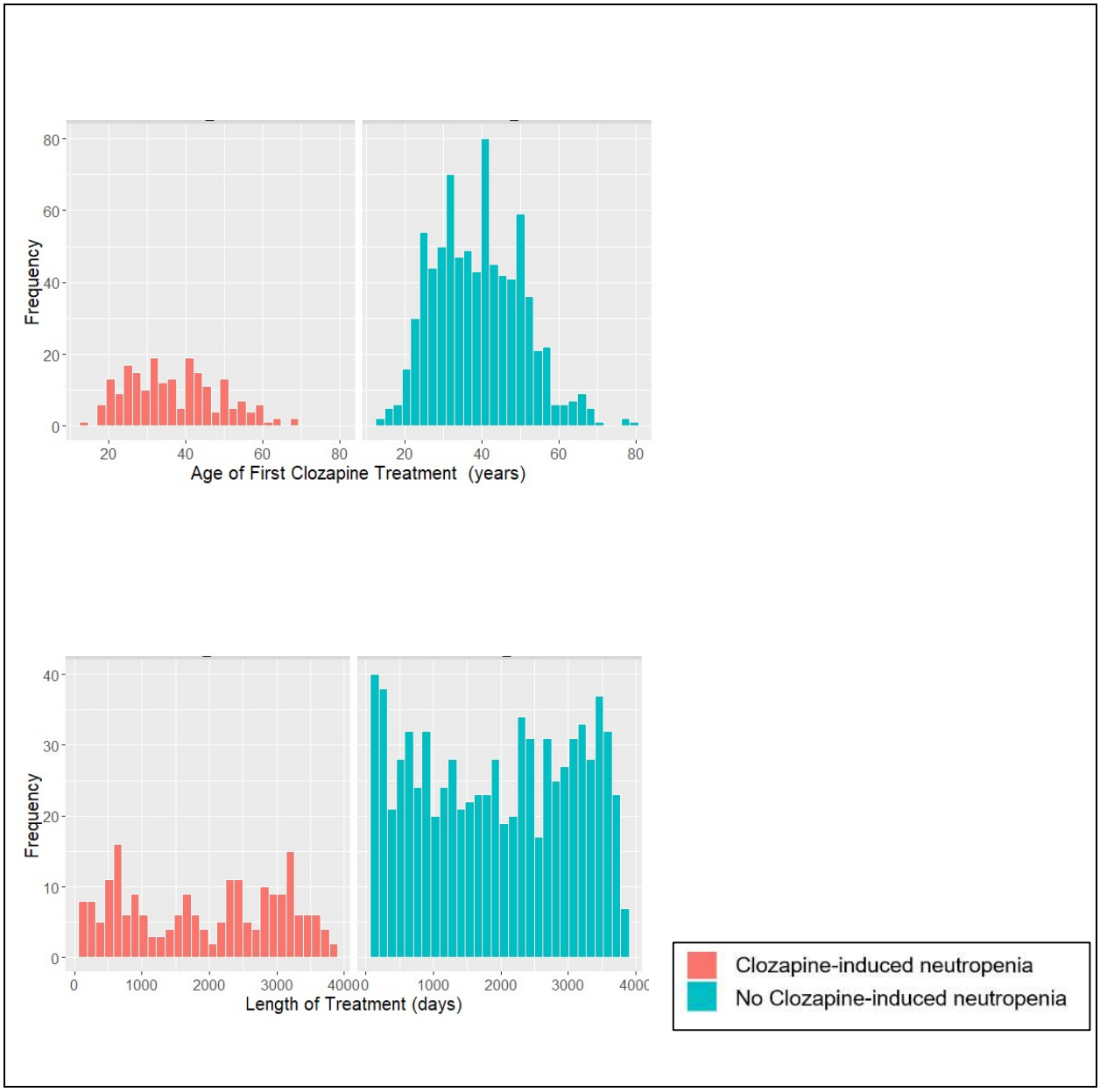
The odd ratio for black ethnicity (OR=2.48, 95% CI 1.89-3.27) indicates that black ethnicity is associated with an increase in the probability of clozapine-induced neutropenia, compared to the white ethnicity. The odds ratio for smokers (OR=1.49, 95% CI 1.13-1.96) indicates that smoking during clozapine treatment is associated with an increase in the probability of clozapine-induced neutropenia, compared to not smoking during clozapine treatment.

Table 3.c: Summary statistics of data used in the modelling of clozapine-induced neutropenia.

Variables	**Summary Statistics**	**Data Source**
Explanatory variables (continuous)		
Age at first clozapine treatment	Range: (10-82) Mean, SD: 39 ± 12.08	Database Table 1
Length of clozapine Treatment (Days)	Range: (90-3803) Mean, SD: 1592.5 ± 1117.65	Database Table 1
Explanatory variables (categorical)		
Gender	Males: 1167 Females: 593	Database Table 1
Ethnicity	White: 821 Black: 703 Asian: 93 Other: 143	Database Table 1
Smoker*	Yes:1039	Database Table 3
Inpatient*	Yes:608	Database Table 4
Welfare Benefits	Yes: 614	Database Table 5
Outcome Measure		
Clozapine-induced Neutropenia	Neutropenia:295 Severe Neutropenia: 33	Database Table 2

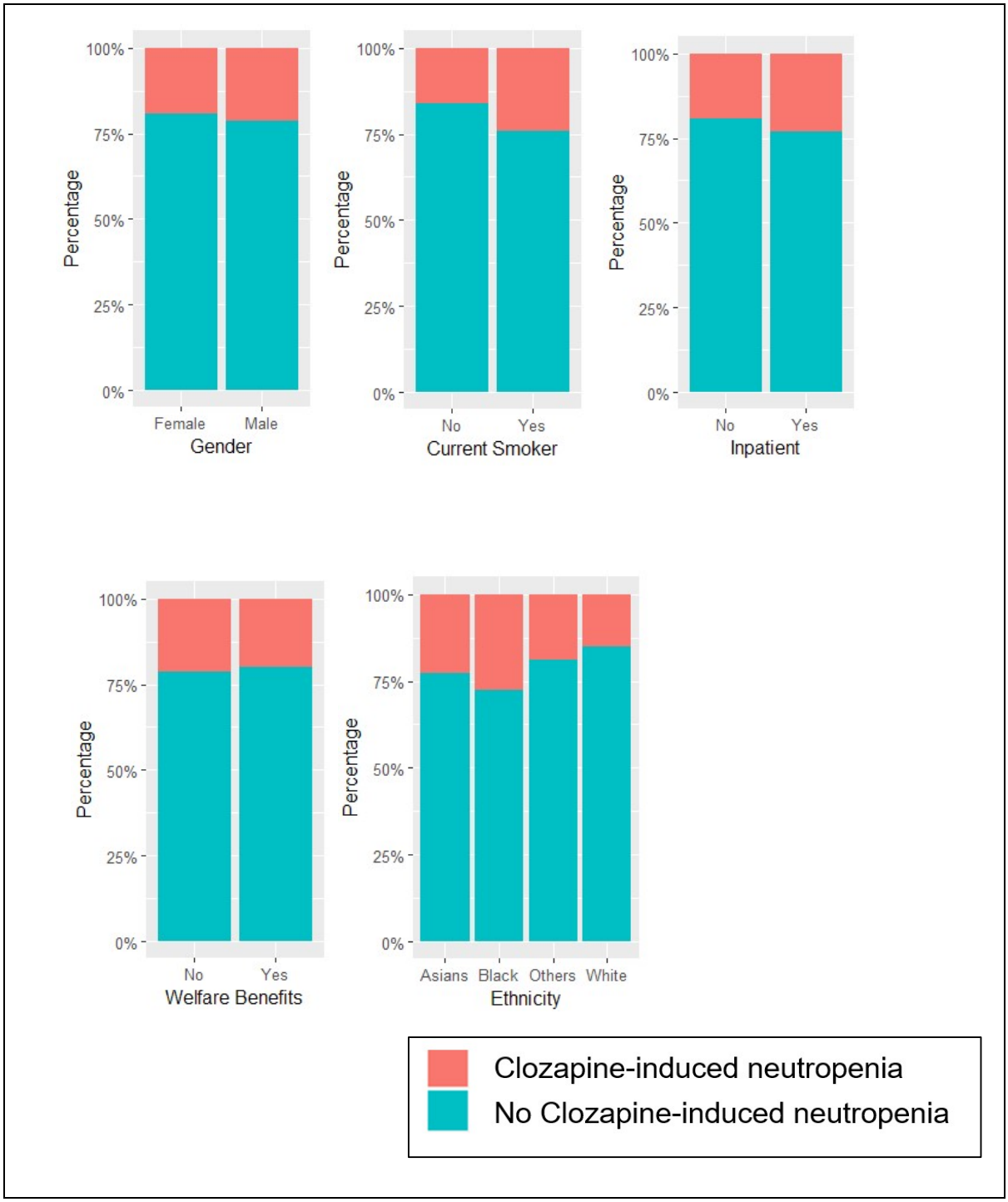
*: during clozapine treatment

Figure 3.b: Histograms of continuous explanatory variables.



The charts show the distribution of patients who had clozapine-induced neutropenia compared to those who did not within each continuous variable.

Figure 3.c: Bar charts of categorical explanatory variables.



The charts above show the proportion of patients who had clozapine-induced neutropenia compared to those who did not within each categorical variable.

Table 3.d: Characteristics of explanatory variables based on clozapine-induced neutropenia status.

	Clozapine-induced neutropenia (n=328)	No clozapine-induced neutropenia (n=1432)	P-value	
Continuous variables (Mann-Whitney U test)				
Age at first clozapine treatment	36.5 ± 11.9	40.0 ± 12.0	3.7e-06	**
Length of Treatment (Days)	1622.9 ± 1101.9	1585.5 ± 1121.5	0.47	
Categorical variables (Fisher exact test)				
Males	234 (71.3%)	933 (65.2%)	0.03	
Ethnicity				
Black	184 (52.6%)	519 (36.2%)	1.5e-04	**
White	99 (33.5%)	722 (50.5%)		
Asian	18 (5.7%)	75 (5.2%)		
Others	27 (8.1%)	116 (8.1%)		
Smoker*	224 (68.3%)	815 (56.9%)	1.4e-04	**
Inpatient*	141 (43.0%)	474 (33.1%)	8.3e-04	**
Welfare Benefits	120 (36.6%)	494 (34.5%)	0.48	

In the table above, the values are mean ± SD or n (%). For calculating significance, Fisher exact test was performed on categorical variables and Mann-Whitney U test was performed on continuous variables Length of Treatment is in years.

Table 3.e: Logistic Regression Results

	Odds ratio	95% CI	P-value	
Continuous variables				
Age at first clozapine treatment	0.98	0.97 - 0.99	6.98e-04	***
Length of Treatment (Days)	1	1 - 1	0.30	
Categorical variables				
Gender				
Female (reference)	1			
Male	1.26	0.96 - 1.66	0.10	.
Ethnicity				
White (reference)	1			
Asian	1.67	0.95 - 2.94	0.07	.
Black	2.48	1.89 - 3.27	8.72e-11	***
Other ethnicity	1.57	0.98 - 2.53	0.06	.
Smoker status*				
Non-smoker (reference)	1			
Smoker	1.49	1.13 - 1.96	4.52e-03	**
Inpatient status*				
Not inpatient (reference)	1			
Inpatient	1.3	0.99 - 1.71	0.06	.
Welfare benefits status				
Not on welfare benefits (reference)	1			
On welfare benefits	0.97	0.77 - 1.26	0.82	

*: during clozapine treatment

3.5 DISCUSSION

3.5.1 Summary of findings

I investigated variables that may be associated with increased risk of neutropenia in clozapine-treated patients so that I can build a predictive model for clozapine-induced neutropenia. I found some potential predictors; however, this current dataset is found to be not informative enough to build a full predictive model for clozapine-induced neutropenia.

In order to predict clozapine-induced neutropenia, I first needed a dataset that had clozapine treatment dates and dates of the neutropenia events. CLZ-ADR algorithm was the only data source in CRIS that provided both this information. Unfortunately, the CLZ-ADR algorithm only reports clozapine treatment that was 90 days or longer. This filtering threshold was pre-set by the algorithm to reduce false positives. It is common for text-mining algorithms to be accompanied with filtering thresholds to implement to reduce false positives – this is a major barrier for using results from text-mining algorithms as that make the results less useful for answering some research questions.

The majority of the clozapine-induced severe neutropenia cases, which is also known as agranulocytosis, are expected to occur in the first few months of clozapine treatment. Amsler et al in 1977 reported that all their observed cases of agranulocytosis occurred during the first 3 months of clozapine (Amsler *et al.*, 1977). Because of this 90-day filtering threshold set by the algorithm, the results from the CLZ-ADR algorithm are not informative enough to be used for building a predictive model for clozapine-induced neutropenia. An alternative source of data is needed.

3.5.2 Alternative source of data

Zaponex Treatment Access System (ZTAS) is one of the UK's mandatory blood monitoring service providers. Clozapine patients treated at South London and Maudsley NHS Foundation Trust (SLAM) have their blood counts monitored by ZTAS (http://www.ztas.co.uk). Additional permissions will be required to access the blood monitoring results of SLAM patients but with the right permissions, this data can be accessed via the CRIS database and there are almost 20 years of data, ranging from May 2000 to October 2019. There are over 200,000 blood monitoring results from over 2,000 SLAM patients. The monitoring results contains the absolute neutrophil counts so it will be straightforward to identify neutropenia events. ZTAS monitoring is only performed on clozapine patients so it will also be straightforward to identify the cohort. I was unable to use ZTAS data for this study because ZTAS data is an external resource with restricted access.

3.5.3 Strengths

As a strength of this study, SLAM is a near-monopoly service provider of all aspects of secondary mental health care to residents within a defined geographic catchment. CRIS being a deidentified version of SLAM's EHR data provides the platform to access information on a range of potential variables that are associated with clozapine-induced neutropenia.

3.5.4 Limitations

Using CRIS data in this study highlighted several challenges around using EHR data effectively in research studies. The following seven challenges summarise typical issues in CRIS and other EHR resources, with reference to the study performed in this thesis chapter.

Challenge 1 – EHR is a vast resource with minimal documentation

EHR is a vast resource, typically comprised of hundreds of database tables, thousands of columns and millions of datapoints. Also, although EHR is a powerful resource for clinical research, it was never built for research purposes and therefore the data is not organised in a way that is easy for researchers to work with. EHR was purposefully built to keep records of patients' health during treatment.

EHR data is so vast that sometimes it is not possible to know about all the different variables that can be extracted from it. EHR data typically comes with minimal documentation. A key factor for designing a research study is to choose the right variables relevant to the study. This step is particularly time-consuming for researchers. This was my first project using CRIS. In order to select the variables for my study, I had to perform exploratory studies on all variables available in database table 5 (demographic table) first. **Table 3.a** shows the different variables that were available in database table 5 and the challenges 2-7 below shows the variable-level challenges present in database table 5, challenges that are typical of EHR data.

This limitation of working with EHR data can be resolved with time, as researchers gain more experience with working with an EHR as prior knowledge of the variables is needed to design a robust study.

Challenge 2 – EHR was not built for research, thorough data standardisation is required

EHR was never built for research purposes, it was built for the primary purpose to keep records of patient health. Because of this, there is little standardisation in EHR data. In addition to the lack of standardisation in the free-text clinical notes that the clinicians write, the lack of standardisation is also present in the structured data.

Data standardisation refers to making the data consistent and improving the quality of the data. One example of lack of standardisation is shown in database table 5 (demographics data), where several values are used to indicate missing data for example, “NULL”, “Not Known”, “Not known”, “Unknown”, “Not Applicable”, “Other”, “Not Disclosed”. Before researchers do any analysis using EHR data, there is usually a time-consuming step of standardising the data by recoding values.

Another example of lack of standardisation is shown by comparing the organisation of data in database table 3 (smoking data) and database table 4 (inpatient data). Database table 3 stores the information on whether the patient smoked during their clozapine treatment. Database table 4 stores the information on whether the patient was an inpatient during their clozapine treatment. One would expect the two database tables to be structured similarly but they are not. The information of non-smokers during clozapine treatment is explicitly provided in database table 3 with a designated category “non-smoker”. The information of patients who were not inpatients during their clozapine treatment is implicit in database table 4. This table only keeps records of dates of when a patient was inpatient, therefore no information is an indication of “not inpatient” and not “missing data”. Prior

knowledge of the assumptions and how missing data is presented in each table is essential to working with EHR data.

Challenge 3 – missing data

Missing data is common in research. However, EHR is known for exceptionally high proportions of missing data in the structured fields. This is the key driver for finding NLP solutions that can help us extract information from free-text fields. In this study, 9 variables were not included in the statistical analysis because they contained more than 25% missing data. We used this conservative threshold for missing data because the statistical analysis involved logistic regression, which excludes from analysis all patients with any missing values. Therefore, instead of losing more patient due to missing data, we made the call to use a conservative threshold for the variables with missing data.

The majority of variables that were excluded due to missing data contained a significantly high proportion of missing data. The percentages of missing data in these variables were: First Language (29%), Employment (44%), Country Of Origin (54%), Lives With (64%), Housing Status (68%), Religion (70%), Has A Twin (77%), Residence (79%) and Asylum Seeker (82%). All the 9 variables came from database table 5 (demographics data), which only stores information that was collected during the patient's registration process at SLAM. Thus, we assumed that these variables were 'missing completely at random' forming a random subset of the participants, and the presence of missing data was not correlated with any other variable. We used the simplest approach to handle missing data, which was to exclude variables with a high proportion of missing data in the statistical analysis, and to analyse individuals who had no missing data at any included variables.

More sophisticated ways to handle missing data exist when more complex patterns of missingness exist in the data. For example, 'missing at random' occurs when the variables are not missing randomly, but their presence or absence is correlated with another variable present in the data set. More complex methods to account for this pattern of missing data are inverse-probability weighting (IPW) and multiple imputation (MI) (Perkins *et al.*, 2018). Both the approaches involve identifying factors that lead to the missingness and then using them to formulate corrected estimates for the missing values. IPW involves each individual receiving a weight representing the probability of them having a missing value (Seaman and White, 2013). Thus, individuals who are most likely to have a missing value are assigned the highest weights. MI involves simulating an unbiased estimate and replacing each missing values with plausible estimates, thus creating complete datasets (Harel *et al.*, 2018). Unfortunately, the strengths of both these methods lies in correctly identifying factors that lead to the missingness (Mansournia and Altman, 2016). This is particularly difficult in EHR research because of the complex interplay of the heterogeneous decisions made by patients and the healthcare providers, thus making approaching missing data an on-going challenge in research based on EHR data (Peskoe *et al.*, 2021).

Challenge 4 – duplicate data

Duplicate variables refer to the same data stored in multiple places. Gender information exists in database table 1 and database table 5. Quality checks showed that that the 2 columns were identical therefore one of the columns was removed.

Ethnicity was another variable that existed in database table 1 and database table 5. Unlike the gender variable, the two columns were not identical, but quality checks showed that

there was consistency between the two. The ethnicity data in database table 5 had 17 categories. Database table 1 had the same data collapsed into 4 categories.

It is typical of EHR data to be stored in multiple places, and the data captured may not always be identical. One needs prior knowledge of the data available in the EHR to identify the most robust resource for each variable.

Challenge 5 – class imbalance

Class imbalance refers to when in a categorical variable, an exceptionally high proportion of values belong to one category. This is another known feature of EHR data. From database table 5 (demographics data), 6 categorical variables had more than 80% data in one category.

Challenge 6 – data in structured fields is not always recent

The data in database table 5 was recorded on the day the patient was admitted. Some patients were admitted over 10 years ago, thus making the variables extracted from this database source not recent. The information like date of birth, gender and ethnicity do not change with time. However, the welfare benefits information that was used from this table is likely to change over time.

This is the downside of using structured data from EHR. Most of the structured data are readily available for research but they are captured at a specific point in the patient's treatment and are not updated. The most recent data on the patients are in the free-text fields, the clinical notes written by clinicians or nurses when they consult with the patients.

Challenge 7 – most of the important information is hidden in the free-text fields

Structured data available in EHR data can be incomplete and full of missing data (challenge 3) or out of date (challenge 6). The most informative part of the EHR data is the free-text fields which contain longitudinal information of the patient's health with in-depth details of important health events. The free-text data are also robust and recent. Unfortunately, breakthroughs in computational techniques are required to harness that true power of data hidden in the free-text formats. Continuous efforts are being made to do this, including in the CRIS data, for example, custom-built Natural Language Processing (NLP) algorithms are used to extract information from the free-text fields, the specifications and performance metrics of which are detailed in an open online catalogue (CRIS NLP Applications Library, 2020).

3.6 CONCLUSION

In conclusion, there are major challenges in working with EHR data which impacted the ability of this study to achieve its aim of identifying predictors of clozapine-induced neutropenia in SLAM patients. One minor oversight during study design, especially when working with text-mined data can lead to irrecoverable limitations in the analysis. This issue arose in this study through using the CLZ-ADR algorithm which omitted the first 90 days of clozapine use, when clozapine-induced neutropenia is more likely to occur. That said, with a cautious study design and thorough awareness of typical EHR challenges, EHR is an extremely useful resource for research.

CHAPTER 4

4 PREPRINT: FREQUENCY OF NEUTROPENIA OVER TIME IN PATIENTS ON CLOZAPINE

This chapter incorporates a paper that has been submitted to medRxiv

Contribution Statement

Amelia Jewell was responsible for creating the data linkage between the ZTAS data and the CRIS data. Upon the completion of the data linkage procedure, James MacCabe and I conceived and designed this study which is based on the linkage data. I performed all the data extraction and statistical analysis as well as wrote the manuscript. All authors critically reviewed the manuscript and approving the final version.

Frequency of neutropenia over time in patients on clozapine

Risha Govind MSc[1,2], Amelia Jewell[2], Eromona Whiskey[2], Siobhan Gee[2], Ebenezer Oloyede[1], David Taylor[2], James H. MacCabe PhD FRCPsych[1,2,3]

1. Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

2. South London and Maudsley NHS Foundation Trust, London, UK.

3. National Psychosis Unit, South London and Maudsley NHS Foundation Trust, London, UK.

Corresponding author: James H. MacCabe (james.maccabe@kcl.ac.uk)

Role of Funding. This work was supported by the Clinical Records Interactive Search (CRIS) system funded and developed by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London and a joint infrastructure grant from Guy's and St Thomas' Charity and the Maudsley Charity (grant number BRC-2011-10035). All authors receive salary support from the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed in this paper are those of the author(s) and not necessarily those of the NIHR, the NHS, or the Department of Health. The above funding had no role in the study design; in analysis, and interpretation of the data; in the writing of the report; and in the decision to submit the paper for publication.

ABSTRACT

Background

Clozapine, the only evidence-based drug for treatment-resistant schizophrenia is associated with agranulocytosis. For this reason, all clozapine patients are required to undergo mandatory regular blood monitoring throughout their clozapine treatment. The blood test results are reported using a traffic light system. The clozapine treatment is stopped immediately after a confirmed red result, which is the indication for risk of agranulocytosis. The need for blood tests places a burden on patients and acts as a barrier to clozapine treatment. There is growing evidence that the risk of agranulocytosis falls steeply after the first few months of treatment, raising the possibility that clozapine monitoring could be discontinued after a certain period of treatment.

Aim

To investigate the frequency density of the confirmed red results from clozapine monitoring across clozapine treatment.

Method

By merging electronic health records (EHR) data with clozapine blood monitoring data, we identified the clozapine treatment dates. The EHR data was from South London and Maudsley NHS Foundation Trust (SLAM). The clozapine blood monitoring data was from Zaponex Treatment Access System (ZTAS). ZTAS is one of the mandatory blood monitoring service providers in the United Kingdom. From these data, Kaplan-Meier survival curve was fitted to determine the time to get confirmed red results. At fixed points in the treatment, the future risk of obtaining a red result were calculated.

Chapter 4
Preprint: Frequency of neutropenia over time in patients on clozapine

Results

By merging over 301,000 data points that came from the blood monitoring results and EHR data of 1,362 patients, we identified 1,891 clozapine treatment periods. Of these, 75 treatments were stopped due to confirmed red results. The Kaplan-Meier survival curve and the incidence rates data showed that 56 (74.7%) confirmed red results occur within the first 6 months of clozapine treatment.

Conclusion

We found a contrast between the relatively high density of the confirmed red results at the beginning of clozapine treatment which significantly reduces after 6 months of treatment which remained low thereafter.

Conflict of interest. JHM has received research funding from Lundbeck.

Ethics statement. The research was conducted under ethical approval reference 18/SC/0372 from Oxfordshire Research Ethics Committee C.

INTRODUCTION

Clozapine, the only evidence-based medication for treatment-resistant schizophrenia, is severely under-prescribed (1–3). This underutilization of clozapine is largely attributed to concerns regarding the risk of clozapine-induced agranulocytosis, an adverse drug reaction of clozapine that is prevalent in 0.4% of patients on clozapine (4–6).

Clozapine-induced agranulocytosis is a rare event with an unknown aetiology (7,8). Agranulocytosis, also known as severe neutropenia, is characterised as extremely low neutrophil count that results in increased susceptibility to fatal infections. Agranulocytosis is defined as a neutrophil count of $<0.5x10^9$/L and neutropenia is defined as a neutrophil count of $<1.5x10^9$/L. Clozapine, an atypical antipsychotic drug, was first introduced in Europe in 1971 for treating patients with schizophrenia (9). A few years later, clozapine was removed from the market after its use was shown to be associated with agranulocytosis (10). A seminal study by Kane et al in 1988 demonstrated its superior efficacy in treatment-resistant schizophrenia and led to the reintroduction of clozapine (1). However, in the United Kingdom (UK), United States and many other nations, clozapine use is subject to mandatory full blood count monitoring for the entire duration of clozapine treatment.

Under the current monitoring regulations in the UK, the blood monitoring starts with a baseline test and the frequency of the blood monitoring decreases with the length of clozapine treatment. A 'baseline test' refers to the monitoring test performed before clozapine treatment is started. The purpose of the baseline test is to ensure that the blood test results are stable before clozapine treatment can be allowed to be initiated. Once clozapine treatment is started, the patient is required to have their full blood counts monitored every week for 18 weeks, as shown in Box 1 (11). If the blood counts are stable in

this period, then from 18 to 52 weeks, the blood monitoring is reduced to fortnightly. If the blood counts continue to be stable after 52 weeks of clozapine treatment, then blood monitoring is reduced to 4-weekly, but cannot be discontinued unless the patient stops taking clozapine (12). In the UK, the white blood counts (WBC) and absolute neutrophil count (ANC) are used to classify the results as either green, amber, or red (13). The patient's further management is guided by this classification (Box 2).

Clozapine treatment must be stopped after a confirmed red result. This means that in the blood monitoring process, as soon as a red result is reported, a follow-up blood test is arranged. The clozapine treatment is stopped immediately if the follow-up blood test results is also red, thus confirming the initial red result. In the UK and Ireland, there is a 'Central Non-Rechallenge Database' to register patients who have had confirmed red results (14). The ANC for red results is $< 1.5x10^9/L$, which is the definition for neutropenia, therefore red results can be used to indicate neutropenia.

The mandatory blood monitoring throughout the clozapine treatment is a major burden to patients and health services, and contributes to the underutilisation of clozapine (15). In this study, we studied the frequency density of the confirmed red results over time using the results from the mandatory blood monitoring.

Box 1: UK clozapine monitoring frequency

Duration of treatment	Monitoring frequency
First 18 weeks	Weekly
19-52 weeks	Fortnightly
>52 weeks	4-Weekly

Box 2: UK clozapine monitoring results classification criteria

Classification	WBC (x 10^9/L)	ANC (x 10^9/L)	Guidance
● Green	>3.5	>2.0	Continue treatment
● Amber	3.0-3.5	1.5-2.0	Continue, but monitor twice weekly until green results
● Red*	<3.0	<1.5*	Arrange emergency blood test to confirm red result. If red result is confirmed, then STOP TREATMENT

**: ANC < 1.5x10^9/L is also the definition of neutropenia. Therefore, red results can be used indicate neutropenia.*

METHOD

Data sources

CRIS

The Clinical Record Interactive Search (CRIS) is a database containing the fully de-identified health records of SLAM. SLAM caters to all secondary mental health care needs of over 1.3 million people of four London boroughs (Lambeth, Southwark, Lewisham, and Croydon); CRIS provides the platform for all their electronic health records to become available to researchers for secondary analysis within a robust data security and governance framework (16).

CRIS is comprised of both structured and free-text fields from the SLAM's clinical notes. In this study, two types of data were extracted from CRIS, pharmacy dispensary data and clozapine clinic attendance data.

The pharmacy dispensary data included the strength and quantity of clozapine that was dispensed to the patient. This data was available in the structured fields. The earliest record of pharmacy data used in this study was from September 2005.

The clozapine clinic attendance data came from the nurse's record of the patient attending the appointment for their routine clozapine blood test. This information was retrieved by combining information from free-text fields as well as structured fields. The free-text fields were searched for the phrases "clozapine clinic" or "clozaril clinic". The structured fields were searched for the entry "attended". The dates on which both the components were retrieved became the clozapine clinic attendance data. The two phrases, "clozapine clinic" or "clozaril clinic" were selected based on exploring the free-text notes and finding that even though the brand Zaponex is used at SLAM, some healthcare providers who write clinical notes tend to refer to refer to clozapine using the brand name 'clozaril'. The earliest record of clozapine clinic attendance data used in this study was from November 2002.

ZTAS

Zaponex Treatment Access System (ZTAS) is one of the UK's mandatory blood monitoring service providers. Clozapine patients treated at South London and Maudsley NHS Foundation Trust (SLAM) have their blood counts monitored by ZTAS (http://www.ztas.co.uk).

The ZTAS data was available via the SQL Server Management Studio version 15.0 (Microsoft Inc, USA). In three separate SQL database tables, ZTAS stores detailed information on the clozapine blood monitoring results (described further below), detailed information on clozapine treatment statuses (described further below) that were recorded, and clozapine treatment start dates (described further below.

The blood monitoring SQL database table of ZTAS includes the date, white blood counts (WBC), absolute neutrophil counts (ANC) and the result classification of each blood test records. It also includes labels for blood tests that were a baseline. Box 2 gives details of how the results are classified. The classifications of the results are either green, amber and red.

Green results means continue treatment, amber means increase monitoring and red means to reconfirm this result, and a confirmed red result requires the immediate discontinuation of clozapine treatment.

The treatment status SQL database table of ZTAS includes the date and status label for all recorded updates in the status of clozapine treatments. The status labels can change multiple times for the same patient. Examples of the statuses are 'on-treatment', 'interrupted', 'discontinued', 'transferred' and 'non-rechallengable'.

The start status SQL database table of ZTAS contained only one clozapine treatment start date per patient. This date was the closest treatment start date prior to the date the linkage was made to access ZTAS data via the CRIS platform (described further below).

All data from these three SQL database tables that were accessible via CRIS were used in this study.

Linkage and cohort definition

A linkage was made from the CRIS database to the ZTAS database. The linkage was performed by mapping patient identifiers from CRIS (including name, NHS number, date for birth) to those in the ZTAS data, and then pseudo-anonymising the data, thus making it seamless to perform secondary analysis using the ZTAS data while patient identities remain de-identified (17,18).

The databases were initially linked in March 2016 and the linkage was updated in October 2019, therefore even though ZTAS continues to monitor SLAM patients, our dataset includes only those who were on both databases on either or both of the linkage dates. This comprised of almost 20 years of ZTAS data, ranging from 2nd May 2000 to 1st October 2019, inclusive.

All CRIS data were extracted until the date of the last record available from ZTAS, 1st October 2019. The cohort comprised all SLAM patients who were in the ZTAS linkage database in CRIS.

Identification of clozapine treatment start dates

SQL Server Management Studio version 15.0 (Microsoft Inc, USA) was used to extract the data and standard Python (V.3.7.4) libraries were used for identifying the clozapine treatment dates.

The clozapine treatment dates that were provided by ZTAS only included one start date per person, and this was the closest start date prior to the date of linkage with CRIS data. In order to find all the treatment start dates, we developed an algorithm using data from ZTAS and CRIS. The data from ZTAS included clozapine blood monitoring results, clozapine treatment statuses and clozapine treatment start dates. The data from CRIS included the pharmacy dispensary data and the clozapine clinic attendance data. Figure 1 shows the study flow chart.

Data Cleaning

The first step of data cleaning was to remove orphan blood tests from the analysis. A baseline test is mandatory before the clozapine treatment is started. Therefore, if a single test exists with no subsequent blood tests, then it can be assumed that that the clozapine treatment was not started at that time. We defined an orphan blood test as a single blood test with no blood tests within 60 days, prior or subsequent to it. The 60 days threshold was used to accommodate for 1 missing datapoint since most of the tests are at 4-weekly frequency.

The second step of data cleaning was to remove the blood test data that occurred immediately after the clozapine treatment was stopped due to a confirmed red result, so that only blood tests on clozapine treatment were included. This criteria was created on the basis

that clozapine treatments are stopped as soon a confirmed red result occurs. It is common for the blood monitoring to continue after the clozapine treatment is stopped due to confirmed red results until the patient's blood results are back in the green zone (Box 2).

Probable treatment dates

Using the cleaned data, we identified the probable start and end dates of clozapine treatment.

A probable start date of clozapine treatment was identified if any of these 3 criteria were fulfilled:

(1) The blood test was the first blood test result of a patient in our dataset

(2) The presence of a 'baseline' label on blood test result

(3) The blood test was after a gap of >35 days for short treatments (see below for definition) or a gap of >60 days for long treatments (see below).

A treatment was classified as 'short' if it was up to 18 weeks long, and 'long' if it was longer than 18 weeks.

A probable end date of clozapine treatment was identified if any of these 4 criteria were fulfilled:

(1) The blood test was the last blood test result of a patient in our dataset

(2) The presence of a confirmed red result

(3) The blood test occurred immediately prior to a gap of >35 days for short treatments or >60 days for long treatments

(4) The ZTAS status changed to 'discontinued', 'interrupted' or 'transferred'.

Once the probable treatment dates were identified, we performed an additional quality control check by examining the frequency of blood tests around each probable start date and end date. Any probable start dates, and probable end dates appearing in the middle of an on-going treatment period and followed by tests at a frequency other than weekly, were removed and not used in the analysis.

Treatments with identifiable start dates

Where clozapine treatment was started when the patient was under the care of a different healthcare provider, and the patient was later transferred to SLAM, we had missing data regarding the start date of the clozapine treatment. For these treatments, the probable start dates that we identified were pointing to the time the clozapine treatment started in SLAM instead of the actual start of the clozapine treatment. Therefore, treatments without identifiable start dates were removed. Since it is mandatory for all clozapine treatments to start with a baseline test, followed by weekly blood tests for the first 18 weeks, a treatment had an identifiable start date if the first test of the treatment had a baseline label AND the mode or mean of the frequency of tests in the first 18 weeks was ≤ 7 days.

To cater for the possibility of missing baseline labels, two additional requirements needed to be fulfilled if the first test of the treatment did not have a baseline label. The additional requirements for these treatments were that the median or mean of the frequency of the tests between 13th to 18th week of treatment was ≤ 7 days AND if the treatment was longer than 18 weeks, then there were >15 tests in the first 18 weeks of treatment.

To cater for treatments that would come under of category of ‘off-licence’, treatments that started any time after the patient had a confirmed red result and therefore would have been

in the non-rechallengable register needed to fulfil an additional requirement. 'Off-licence' refers to the use of clozapine outside the marketing authorisation issued in the UK for the drug, meaning the patient is receiving clozapine treatment after being registered register as non-rechallengable (19). The additional requirement for these treatments was for their ZTAS status to change from 'non-rechallegable' to 'on-treatment'. Note that these treatments were not comprised of the tests that immediately followed a confirmed red result as those blood tests was already removed earlier in the analysis.

The treatments with unidentifiable start date that ended due to confirmed red results were manually checked to confirm that it was indeed not possible to identify the start dates of these treatments within the scope of the existing data.

Statistical Analysis

The statistical analysis was performed using STATA for Windows version 15.1. A Kaplan-Meier survival curve was fitted to display the time to confirmed red results. The confirmed red results were used as the failure event in this analysis.

The *stptime* command of STATA was used to compute and tabulate the person-years and incidence rates of confirmed red results against the length of clozapine treatment. The person-years is the sum of the number of years each patient has been on treatment. The incidence rates refer to the number of confirmed red results divided by the person-years. Due to the low incidence of confirmed red results in certain period of treatment, the incidence rates were measured per 1,000 person-years.

Using the person-years and number of confirmed red results data, the future incidence rates of red results at different points of clozapine treatment were calculated. First, the future

person-years were calculated by finding the sum of the number of years of treatments remaining at specific point in treatments. These future incidence rates were calculated by dividing the future number of confirmed red results by the future person-years at specific points in clozapine treatment. The future incidence rates were also measured per 1,000 person-years. This calculation was performed using Microsoft Excel version 2102 (Microsoft Inc, USA).

Ethical considerations

CRIS was approved for use as a de-identified data resource for secondary analysis by Oxfordshire Research Ethics Committee C (reference 18/SC/0372).

RESULTS

Between 2nd May 2000 and 1st October 2019, ZTAS recorded 210,273 blood test results for 2,028 SLAM patients. The number of blood tests per person ranged from 1 to 341. The median number of tests per person was 94.

Figure 2 shows that after removing orphan blood tests and blood tests that were performed immediately after the confirmed red result, there were 208,554 tests remaining. These came from 1,988 SLAM patients and comprised of 3,167 probable treatment periods.

Of the 3,167 probable treatments, 1,276 (40%) had an unidentified start date, and were thus excluded from the analysis. 24 (1.9%) of the treatments with unidentifiable start dates ended in confirmed red results. These 24 treatments were manually checked, and this confirmed that all these treatments started outside the scope of the data available.

Of the 3,167 probable treatments, we were able to identify the start dates of 1,891 (60%) of them, of which 1,551 (82%) had their first test labelled as 'baseline'. The treatment lengths ranged from 2 days to 19 years, with a median of 1.1 years. 75 of these treatments ended with the confirmed red results. The majority (74.7%) of the treatments that ended due to confirmed red results ended within the first 6 months of treatment.

Figure 3 shows the distribution of the confirmed red results over time. This plot was generated using the 75 treatments that ended with confirmed results with known start dates. It shows that majority of the confirmed red results occur in the first 6 months of treatment. After 6 months, the incidence of confirmed red results is sporadic. This pattern is reflected in the Kaplan Meier survival analysis and in the incidence rates the future incidence.

Figure 4 shows the Kaplan-Meier survival plot of the length of clozapine treatment to get confirmed red results. This plot was generated using the 1,891 treatments that had known start dates. Of the 75 treatments that ended with confirmed red results, 56 (74.7%) confirmed red results occur in the first 6 months of clozapine treatment. The plot demonstrates 3 distinct phases of risk for getting a confirmed red result: (I) the risk is highest in the first 6 months of treatment (II) the risk reduces to a reasonably constant level from after 6 months to 7 years of treatment (III) after 7 years the risk is almost zero as, after 7 years, there is only 1 incidence of a confirmed red result, and that is at 10.4 years.

Table 1 shows the person-years and incidence rates of confirmed red results at different time points in clozapine treatment. Compared with the rest of the treatment, the first year of treatment has a higher incidence rate of confirmed red results, 47 (95% CI: 36-60) per 1,000 person-years. The incidence rate of confirmed red results in the second year of treatment is 5 (95% CI: 2-12) per 1,000 person-years. The majority of confirmed red results occur in the

first 4 months of clozapine treatment. The overall incidence rate of confirmed red results at any time in the treatment is 14 (95% CI: 11-17) per 1,000 person-years.

Table 2 shows the future person-years and incidence rates of confirmed red results at different time points in clozapine treatment. The rate of confirmed red results are the highest at the beginning of treatment. The future incidence rate of confirmed red results in the 1st month of treatment is 13.8 per 1,000 person-years. This rate gradually decreases until the 6th month of treatment, where the future incidence rate of confirmed red results is 4.0 per 1,000 person-years. The future incidence rate remains below 4.0 for the rest of the treatment.

DISCUSSION

Summary of findings

We investigated if risk of clozapine-induced neutropenia changes with clozapine treatment duration using the clozapine blood monitoring data. We used the confirmed red results as the indication of neutropenia. We found a contrast between the relatively higher risk at the beginning of clozapine treatment and significantly reduced after 6 months of treatment.

Comparison with previous studies

To our knowledge, no previous research has specifically investigated the risk of neutropenia using clozapine blood monitoring data. Our results are consistent with the findings of Amsler et al in 1977 where they reported that all their observed cases of severe neutropenia occurred during the first 3 months of clozapine (20).

Strengths and limitations

As a strength of this study, this study was based on SLAM patients. SLAM has a near-monopoly in providing all aspects of secondary mental health care to over 1.3 million people of four London boroughs (Lambeth, Southwark, Lewisham, and Croydon), thus creating an ascertainment to study the effects of antipsychotics drugs, such as clozapine on patients in the UK.

The main limitation of this study is missing data. The study was performed on data available to us from the electronic health records of SLAM and the clozapine blood monitoring data from ZTAS, latter was used as the primary data for identifying the dates of clozapine treatment. Unfortunately, we could not identify the start dates of 40% of the treatment periods, which comprised of 47% of blood monitoring data. These clozapine treatments started prior to the date of first blood monitoring data we had of these patients. ZTAS became SLAM's blood monitoring service provider in the year 2000. Since we only have access to the blood monitoring results of one clozapine monitoring service, ZTAS, we were not able to identify their start dates of clozapine treatments that started pre-2000. Similarly, we were not able to identify the start dates of patients who transferred from another trust where they started their clozapine treatment. The treatments with unidentifiable start dates were removed from the analysis, and this included 24 treatments that ended with confirmed red results. We manually checked the data on these 24 treatments to confirm that they start dates were unidentifiable using the data available to us.

Another limitation was that although the information on clozapine start dates were embedded in the free-text clinical notes, extracting this information was a major challenge.

Because of the lack of standards in writing the clinical notes, embedded information within the clinical notes is currently not utilized to its fullest potential in research in general (21).

The findings on future incidence of red results should be treated with caution. Some of these results, particularly for later in treatment, are based on low numbers. However, we believe these data should be presented as they make the general point that the future risk of red results (on which mandatory testing is predicated) is substantially lower in patients with long periods of treatment.

Implications

To our knowledge, our study is the first to present data on the risk of having a future confirmed red result and that the risk falls to low levels after 1 year. The risks of a future red result should be weighed against the risks and burdens of monitoring, including the risk of psychotic relapse due to unnecessarily discontinuing clozapine due to a low neutrophil count that is unrelated to clozapine treatment. This supports the case for reducing the monitoring even further as treatment progresses because as treatment progresses, the risk-benefit ratio of monitoring changes significantly. It also suggests that the rigid application discontinuation rules based on thresholds may not be appropriate, at least after 6 months. An alternative system could be proposed, whereby, if a confirmed red result is obtained after 6 months, a haematological review could be triggered to try to determine whether the cause of the neutropenia was likely to be related to clozapine or not, and to advise the treating team on the likely risks of clozapine rechallenge.

Future Work

Since with the current methods, we were only able identify the start dates of just 53% of the treatments, we will be exploring computational solutions such as Natural Language Processing (NLP) to identify more clozapine start dates from information embedded in the free-text clinical notes. Also, we will perform cost-effectiveness analysis to model the tipping point where the benefit of monitoring no longer outweighs the burden and cost of clozapine monitoring to the healthcare system as well as the patient.

Figure 1:Study Flow chart

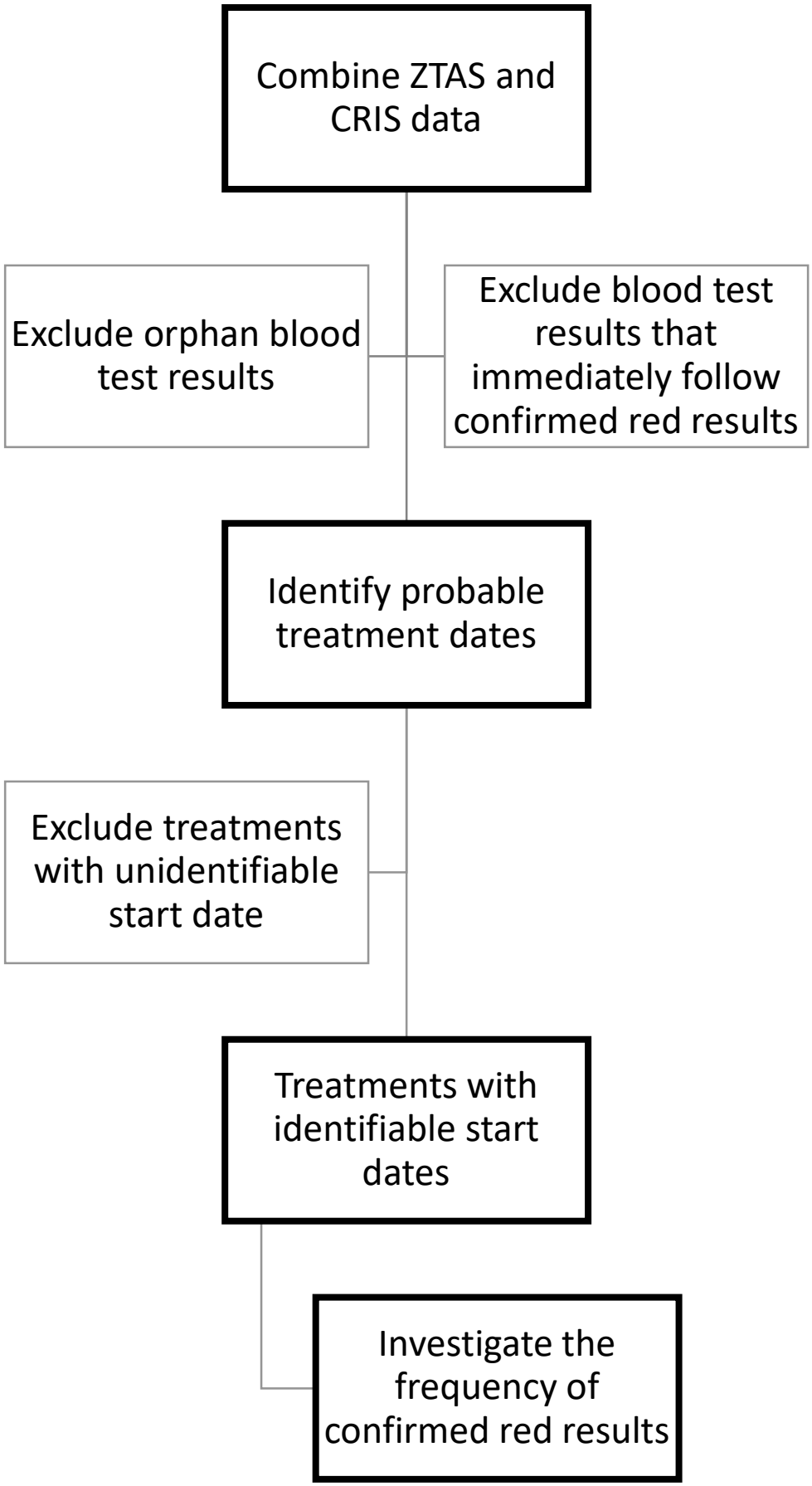


Figure 2: Figure showing the numbers of probable treatment periods identified, of these, the treatment periods with identifiable start dates. Treatment periods were calculated from combining blood test results, dispensary data and clinic attendance data. Not all treatments had identifiable start dates due to missing data, as those treatments started in other settings.

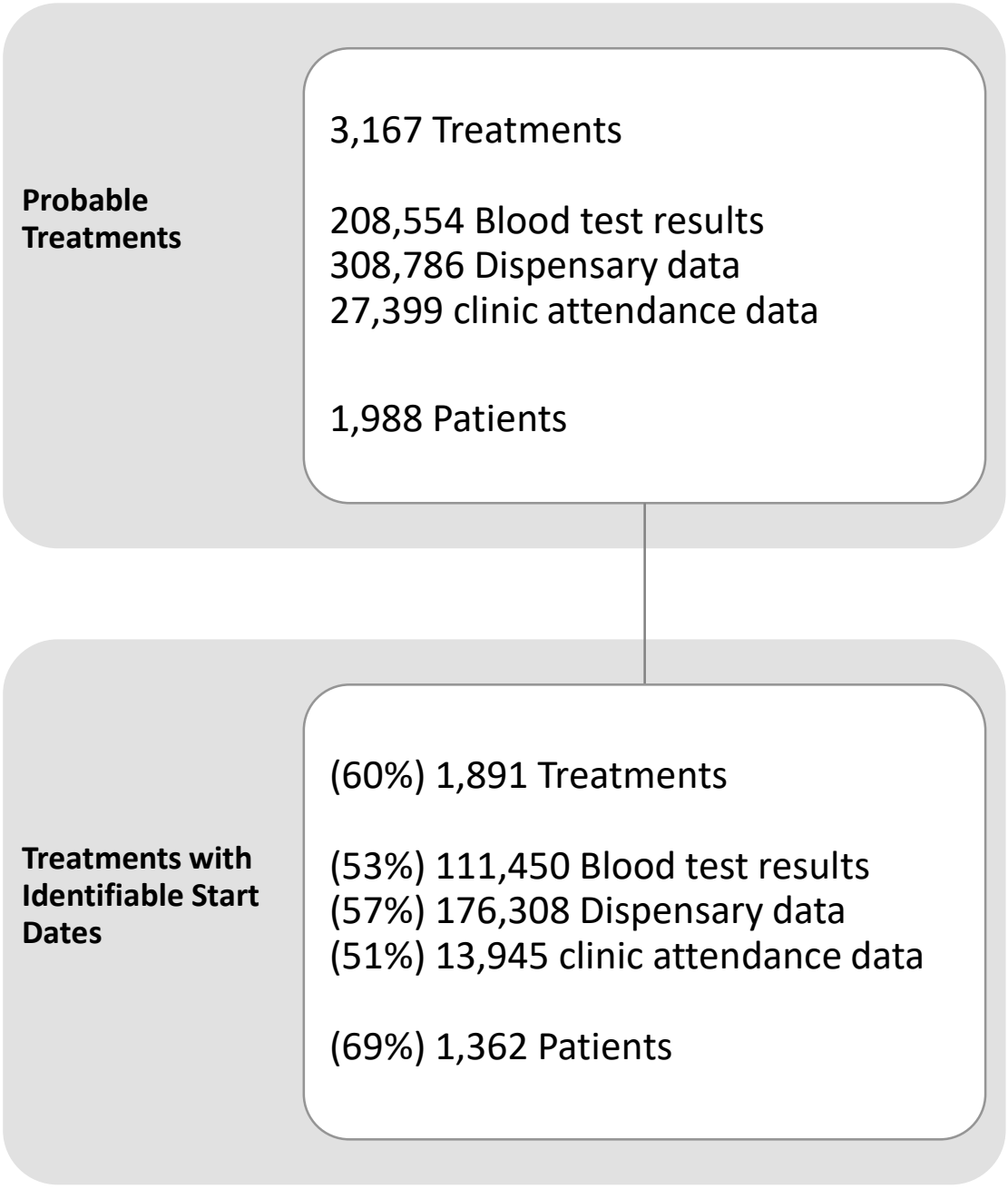


Figure 3: The Histogram of confirmed red results over time plot shows that majority of the confirmed red results occur in the first 6 months. After 6 months of treatment, the incidence of confirmed red results is sporadic. This pattern is reflected in the Kaplan Meier survival analysis and in the incidence rates the future incidence. One explanation of this pattern could be that there are two distinct biological mechanisms, and the clozapine-induced immunological response occurs in the first 6 months of clozapine treatment and afterwards, the occurrence of the confirmed red results is random.

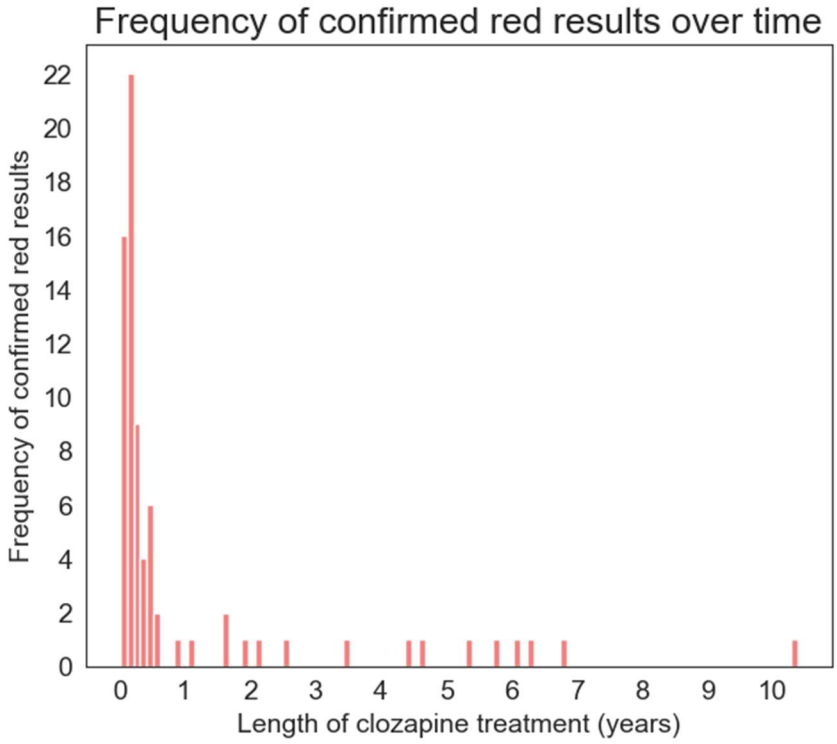
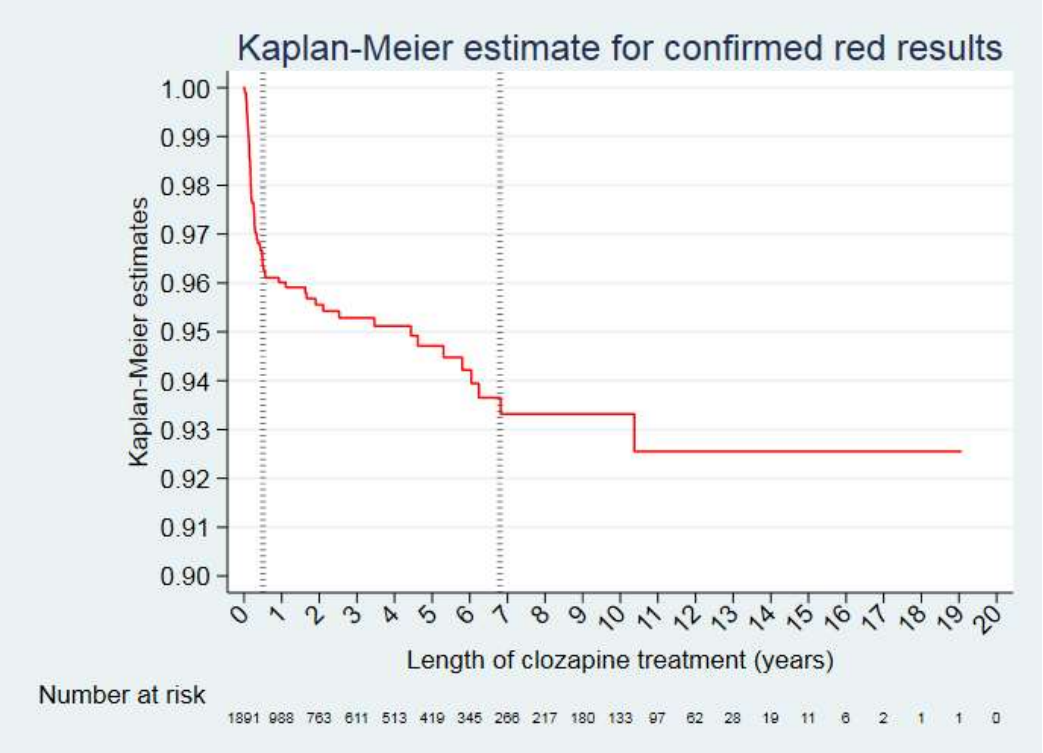


Figure 4: Kaplan-Meier plot of confirmed red results. Below the plot shows the number at risk at each year. The plot demonstrates 3 distinct phases of risk for getting a confirmed red result: (I) the risk is high in the first 6 months of treatment (II) the risk is low and reasonably constant from after 6 months to 7 years of treatment (III) after 7 years the risk is almost zero as, after 7 years, there is only 1 incidence of a confirmed red result



Chapter 4

Preprint: Frequency of neutropenia over time in patients on clozapine

Table 1: Number of confirmed red results, person-years, and incidence rates of the confirmed red results as clozapine treatment progresses, shown for every month for the first 12 months, then for every year. Person-years is the sum of years in all the treatments that ended within the specified period. The incidence rate for each period is refers to the number of confirmed red results divided by the person-years. The incidence rates are displayed per 1,000 person-years.

Length of clozapine treatment	Number of confirmed red results in this period	person-years	Incidence Rate per 1,000 person-years (95% CI)
0-1 years	60	1284.5	47 (36-60)
0-1 months	10	147.1	68 (37-126)
1-2 months	17	132.6	128 (80-206)
2-3 months	11	122.6	90 (50-162)
3-4 months	9	116.3	77 (40-149)
4-5 months	3	110.2	27 (9-84)
5-6 months	6	105.3	57 (26-127)
6-7 months	3	100.1	30 (10-93)
7-8 months	0	96.9	0
8-9 months	0	93.4	0
9-10 months	0	90.0	0
10-11 months	0	86.4	0
11-12 months	1	83.6	12 (2-85)
1-2 years	4	857.1	5 (2-12)
2-3 years	2	683.9	3 (1-12)
3-4 years	1	558.0	2 (0-13)
4-5 years	2	468.2	4 (1-17)
5-6 years	2	385.7	5 (1-21)
6-7 years	3	304.4	10 (3-31)
7-8 years	0	240.4	0
8-9 years	0	197.1	0
9-10 years	0	156.6	0
10-11 years	1	115.9	9 (1-61)
11-12 years	0	78.5	0
12-13 years	0	43.8	0
13-14 years	0	22.1	0
14-15 years	0	14.4	0
15-16 years	0	9.1	0
16-17 years	0	3.9	0
17-18 years	0	1.9	0
18-19 years	0	1.0	0
19-20 years	0	0.1	0
total	75	5426.6	14 (11-17)

Chapter 4

Preprint: Frequency of neutropenia over time in patients on clozapine

Table 2: Table of the incidence rate (per 1,000 person-years) for getting a red result in at different time points in clozapine treatment. This statistic was calculated by dividing the future number of confirmed red results by the future person-years at specific points in clozapine treatment. The future person-years is the sum of the number of years of treatments remaining at specific point in treatments.

Point in clozapine treatment	Future number of confirmed reds	Future person-years	Future Incidence Rate per 1,000 person years
0 months	75	5426.6	13.8
1 month	65	5279.5	12.3
2 months	48	5146.9	9.3
3 months	37	5024.3	7.4
4 months	28	4908.0	5.7
5 months	25	4797.8	5.2
6 months	19	4692.5	4.0
7 months	16	4592.4	3.5
8 months	16	4495.5	3.6
9 months	16	4402.1	3.6
10 months	16	4312.1	3.7
11 months	16	4225.7	3.8
1 year	15	4142.1	3.6
2 years	11	3285.0	3.3
3 years	9	2601.1	3.5
4 years	8	2043.1	3.9
5 years	6	1574.9	3.8
6 years	4	1189.2	3.4
7 years	1	884.8	1.1
8 years	1	644.4	1.6
9 years	1	447.3	2.2
10 years	1	290.7	3.4
11 years	0	174.8	0
12 years	0	96.3	0
13 years	0	52.5	0
14 years	0	30.4	0
15 years	0	16.0	0
16 years	0	6.9	0
17 years	0	3.0	0
18 years	0	1.1	0
19 years	0	0.1	0
20 years	0	0	0

Chapter 4

Preprint: Frequency of neutropenia over time in patients on clozapine

REFERENCES

1. Kane J, Honigfeld G, Singer J, Meltzer H. Clozapine for the Treatment-Resistant Schizophrenic: A Double-blind Comparison With Chlorpromazine. Arch Gen Psychiatry. 1988;45(9):789–96.
2. Howes OD, Vergunst F, Gee S, McGuire P, Kapur S, Taylor D. Adherence to treatment guidelines in clinical practice: Study of antipsychotic treatment prior to clozapine initiation. Br J Psychiatry. 2012 Dec;201(6):481–5.
3. Siskind D, McCartney L, Goldschlager R, Kisely S. Clozapine v . first- and second-generation antipsychotics in treatment-refractory schizophrenia: systematic review and meta-analysis . Br J Psychiatry. 2016 Nov;209(5):385–92.
4. Li XH, Zhong XM, Lu L, Zheng W, Wang S Bin, Rao WW, et al. The prevalence of agranulocytosis and related death in clozapine-treated patients: A comprehensive meta-analysis of observational studies. Psychol Med. 2020 Mar 1;50(4):583–94.
5. Myles N, Myles H, Xia S, Large M, Kisely S, Galletly C, et al. Meta-analysis examining the epidemiology of clozapine-associated neutropenia. Acta Psychiatr Scand. 2018 Aug 1;138(2):101–9.
6. Meltzer HY. Clozapine: Balancing safety with superior antipsychotic efficacy. Clin Schizophr Relat Psychoses. 2012 Oct;6(3):134–44.
7. Mijovic A, MacCabe JH. Clozapine-induced agranulocytosis. Ann Hematol. 2020 Nov 1;99(11):2477–82.
8. Lee J, Takeuchi H, Fervaha G, Powell V, Bhaloo A, Bies R, et al. The Effect of Clozapine on Hematological Indices: A 1-Year Follow-Up Study. J Clin Psychopharmacol. 2015 Oct 12;35(5):510–6.
9. Bablenis E, Weber SS, Wagner RL. Clozapine: A novel antipsychotic agent. DICP, Ann Pharmacother. 1989;23(2):109–15.
10. Idänpään-Heikkilä J, Alhava E, Olkinuora M, Palva IP. Agranulocytosis during treatment with clozapine. Eur J Clin Pharmacol. 1977 May;11(3):193–8.
11. Taylor D, Barnes TRE, Young AH. The Maudsley prescribing guidelines in psychiatry. London: Wiley-Blackwell; 2018.
12. BNF. BNF: British National Formulary - NICE. BMJ Group and Pharmaceutical Press; 2020.
13. Nielsen J, Young C, Ifteni P, Kishimoto T, Xiang YT, Schulte PFJ, et al. Worldwide differences in regulations of clozapine use. Vol. 30, CNS Drugs. Springer International Publishing; 2016. p. 149–61.
14. Oloyede E, Casetta C, Dzahini O, Segev A, Gaughran F, Shergill S, et al. There Is Life After the UK Clozapine Central Non-Rechallenge Database. Schizophr Bull. 2021 Jul 8;47(4):1088–98.
15. Bachmann CJ, Aagaard L, Bernardo M, Brandt L, Cartabia M, Clavenna A, et al. International trends in clozapine use: a study in 17 countries. Acta Psychiatr Scand. 2017 Jul 1;136(1):37–51.
16. Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, et al. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. BMC Psychiatry. 2009 Dec 12;9(1):51.
17. Perera G, Broadbent M, Callard F, Chang C-K, Downs J, Dutta R, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open. 2016;6(3).
18. Jewell A. The Clinical Data Linkage Service (CDLS) [Internet]. NIHR South London and Maudsley NHS Foundation Trust Biomedical Research Centre; 2018. Available from: https://www.maudsleybrc.nihr.ac.uk/media/219616/the-clinical-data-linkage-service-cdls.pdf
19. Meyer N, Gee S, Whiskey E, Taylor D, Mijovic A, Gaughran F, et al. Optimizing outcomes in clozapine rechallenge following neutropenia: A cohort analysis. J Clin Psychiatry. 2015 Nov 1;76(11):e1410–6.
20. Amsler HA, Teerenhovi L, Barth E, Harjula K, Vuopio P. Agranulocytosis in patients treated with clozapine: A STUDY OF THE FINNISH EPIDEMIC. Acta Psychiatr Scand. 1977 Oct 1;56(4):241–8.
21. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc. 2011 Mar;18(2):181.

Chapter 4

Preprint: Frequency of neutropenia over time in patients on clozapine

CHAPTER 5

5 PUBLICATION: CLOZAPINE TREATMENT AND RISK OF COVID-19 INFECTION: RETROSPECTIVE COHORT STUDY

This chapter incorporates the following publication:

> Govind, R. et al. (2020) 'Clozapine treatment and risk of COVID-19 infection: retrospective cohort study', *The British Journal of Psychiatry*. Royal College of Psychiatrists, pp. 1–7. doi: 10.1192/bjp.2020.151.

Contribution Statement

This study was conceived by Richard Hayes and James MacCabe. They are both joint senior authors on the manuscript. Daniela Fonseca de Freitas, Richard Hayes, James MacCabe and I designed the study. With the support and guidance from other authors, I performed all the data extraction and statistical analysis for this chapter; Megan Pritchard supervised the data extraction part; Daniela Fonseca de Freitas supervised the statistical analysis part. I wrote the Methods and Results section of the manuscript. All authors contributed to manuscript preparation and approving the final version.

BJPsych | The British Journal of Psychiatry (2020)
Page 1 of 7. doi: 10.1192/bjp.2020.151

Clozapine treatment and risk of COVID-19 infection: retrospective cohort study

Risha Govind, Daniela Fonseca de Freitas, Megan Pritchard, Richard D. Hayes* and James H. MacCabe*

Background

Clozapine, an antipsychotic with unique efficacy in treatment-resistant psychosis, is associated with increased susceptibility to infection, including pneumonia.

Aims

To investigate associations between clozapine treatment and increased risk of COVID-19 infection in patients with schizophrenia-spectrum disorders who are receiving antipsychotic medications in a geographically defined population in London, UK.

Method

Using information from South London and Maudsley NHS Foundation Trust (SLAM) clinical records, via the Clinical Record Interactive Search system, we identified 6309 individuals who had an ICD-10 diagnosis of schizophrenia-spectrum disorders and were taking antipsychotics at the time of the COVID-19 pandemic onset in the UK. People who were on clozapine treatment were compared with those on any other antipsychotic treatment for risk of contracting COVID-19 between 1 March and 18 May 2020. We tested associations between clozapine treatment and COVID-19 infection, adjusting for gender, age, ethnicity, body mass index (BMI), smoking status and SLAM service use.

Results

Of 6309 participants, 102 tested positive for COVID-19. Individuals who were on clozapine had increased risk of COVID-19 infection compared with those who were on other antipsychotic medication (unadjusted hazard ratio HR = 2.62, 95% CI 1.73–3.96), which was attenuated after adjusting for potential confounders, including clinical contact (adjusted HR = 1.76, 95% CI 1.14–2.72).

Conclusions

These findings provide support for the hypothesis that clozapine treatment is associated with an increased risk of COVID-19 infection. Further research will be needed in other samples to confirm this association. Potential clinical implications are discussed.

Keywords

COVID-19; clozapine; antipsychotics; epidemiology; psychotic disorders.

Clozapine is an antipsychotic with unique efficacy in treatment-resistant psychosis and, for many people, it is the only effective treatment.[1] It is associated with a reduction in hospital admissions, overall mortality and suicide risk in schizophrenia.[2–5] People with schizophrenia have an increased mortality compared with the general population.[6,7] Some of this excess mortality is attributable to pneumonia[8–11] and much of this increase may be attributable to higher rates of smoking.[12] However, there appears to be an additional effect of clozapine treatment.[13–16] In the study of Kuo and colleagues, treatment with clozapine was associated with approximately a doubling of the risk of pneumonia.[14] However, confounding by indication could have affected these results: clozapine is prescribed to people with treatment-resistant schizophrenia, and such individuals are likely to have a range of comorbidities that increase the risk of infection, such as smoking and other substance misuse, poor diet and a sedentary lifestyle.[17] It is also plausible that some of the adverse effects of clozapine, such as diabetes, weight gain and hypersalivation (leading to aspiration pneumonia[18]), could lie on the causal pathway between clozapine treatment and the risk of infection. Clozapine treatment appears to have multiple effects on the innate immune system, including transient eosinophilia, cytokine release and fever during early treatment, and neutropaenia and agranulocytosis in a small minority.[19] There is emerging evidence that adaptive immunity is also affected by clozapine,[20] with a reduction in all three classes of circulating immunoglobulins (IgM, IgA and IgG) in clozapine-treated patients compared with those on other antipsychotics. COVID-19 is a novel infection caused by SARS-Cov-2, causing pneumonia in severe cases. It arose in China in late 2019 and was declared a global pandemic by the World Health Organization (WHO) in March 2020.[21] Given the effects of clozapine on immunity and the increased risk of pneumonia, we investigated whether clozapine treatment was associated with an increased risk of COVID-19 infection in patients with schizophrenia and other psychoses treated with antipsychotics in a geographically defined population in London during the COVID-19 pandemic.

Method

Setting and ethics statement

This retrospective cohort study used data from the South London and Maudsley NHS Foundation Trust (SLAM), one of Europe's largest secondary mental healthcare providers. In the UK, mental health services are provided on the basis of defined geographical catchment areas under the National Health Service (NHS). SLAM provides all aspects of secondary mental healthcare to over 1.3 million people of four London boroughs (Lambeth, Southwark, Lewisham and Croydon). From 2006, SLAM has used a fully electronic health records system and the Clinical Records Interactive Search system (CRIS), supported by the National Institute for Health Research (NIHR) Specialist Biomedical Research Centre for Mental Health. CRIS was established in 2008

* Joint last authors.

to enable researchers to search and retrieve de-identified clinical records from SLAM. The protocol for CRIS has been described in detail in an open-access publication.[22] CRIS was approved as an anonymised data resource for secondary analysis by Oxfordshire Research Ethics Committee C (reference 18/SC/0372). The data linkage to King's College Hospital for admissions regarding COVID-19 infections took place under Regulation 3(2) and Regulation 3(3) of the Health Service Control of Patient Information Regulations 2002 (COPI).

Analytical cohort and data extraction

The cohort comprised individuals who fulfilled all three of the following inclusion criteria: ICD-10 diagnosis of schizophrenia-spectrum disorders (F2*); taking antipsychotic medication between 1 December 2019 and 1 March 2020; and receiving out-patient or in-patient care at SLAM on 1 March 2020. This date was chosen because it was before 12 March 2020, the date of the first diagnosed case of COVID-19 in SLAM, so there was no risk of reverse causation (the presence of COVID-19 infection affecting the exposures).

SQL Server Management Studio version 15.0 for Windows was used to extract the data. The day of data extraction was 18 May 2020. The index period, from which medication data were gathered, was from 1 December 2019 to 1 March 2020. Patients were followed up from 1 March 2020 until they were diagnosed COVID-19 positive, died or reached the end of the observation period (18 May 2020), whichever occurred first.

Even though specific structured fields exist within CRIS, these are often incomplete and much of the useful information in CRIS is within the free-text fields of clinical notes. To fully exploit this database, data from structured fields are augmented by data extracted from free-text fields of clinical records, using custom-built natural language processing (NLP) algorithms.[23] NLP algorithms are able to outperform keyword searches because they take into account the linguistic context of terms of interest, for example temporal modifiers (e.g. 'on clozapine' versus 'previously took clozapine'). Data from four NLP algorithms were used in this study: diagnosis, medication, smoking and body mass index (BMI).

The diagnosis algorithm was used for the inclusion criteria to identify individuals who were ever diagnosed with ICD-10 schizophrenia-spectrum disorders (F2*). The precision and recall scores for the diagnosis algorithm are 100% and 65% respectively.[23]

The medication algorithm data were also used for the inclusion criteria to identify individuals who were on an antipsychotic medication between 1 December 2019 and 1 March 2020, the index period. This algorithm provides specific results for antipsychotic medication. The antipsychotic prescriptions included in this analysis were clozapine, olanzapine, risperidone, aripiprazole, amisulpride, paliperidone, flupentixol, haloperidol, zuclopenthixol, quetiapine, fluphenazine, sulpiride, lurasidone, trifluoperazine, chlorpromazine, pipotiazine, penfluridol, droperidol, pimozide, thioridazine, promazine, ziprasidone hydrochloride, levomepromazine and pericyazine. The precision and recall scores for the antipsychotics part of the medication algorithm are 88% and 90% respectively.[23]

The smoking algorithm was used to identify the smoking status of each patient. The 'current smoker' status was based on data recorded between 1 March 2019 and 1 March 2020. The 'past smoker' and 'never smoked' statuses were based on all available information in the electronic health record. In the underlying patient records, smoking status may be recorded repeatedly, i.e. each time this information is entered into the patient record. Consequently, for some patients, the smoking algorithm may identify more than one smoking status per patient. Where this was the case, we took the highest smoking status in the hierarchy 'current smoker' > 'past smoker' > 'never smoked'. The precision (P) and recall (R) scores for each status of the smoking algorithm are as follows: for 'current smoker' status, $P = 79\%$ and $R = 87\%$; for 'past smoker' status, $P = 68\%$ and $R = 38\%$; for 'never smoked' status, $P = 72\%$ and $R = 75\%$.[23]

The BMI algorithm was used to extract the most recent BMI measurement for each patient in the entire patient record. To exclude erroneous values from the results of this algorithm, we rejected values outside the range 15–70 kg/m^2. The overall precision and recall scores for the BMI algorithm are 89% and 78% respectively.[23]

All the NLP algorithm outputs were also supplemented by the data in the structured fields, data in the health records (such as data from ICD-10 diagnosis forms for diagnosis data) and pharmacy dispensary data for medication data.

Of all patients in SLAM, 6309 met the inclusion criteria of individuals with ICD-10 diagnoses of schizophrenia-spectrum disorders (F2*) who were on antipsychotic medication during the index period.

Main outcome measure

The outcome of interest was infection with COVID-19 during the follow-up period (1 March to 18 May 2020). These data were collated by combining information from the SLAM pathology laboratory results, the presence of a clinician-entered alert on SLAM records reading 'COVID-19 positive' and information provided by local general hospitals (King's College Hospital and Princess Royal University Hospital) for COVID-19-related admissions.

Exposure of interest and potential confounding variables

In keeping with the cohort study design, the exposure of interest and potential confounders were recorded before the start of follow-up. People who were on clozapine treatment at any time between 1 December 2019 and 1 March 2020 were designated as the exposed group. Those on any type or combination of antipsychotic treatment that did not include clozapine during this time constituted the unexposed group.

We considered the following potential confounders: sociodemographic characteristics, health and use of SLAM services. The sociodemographic information was age, gender and ethnicity. The health information was smoking status and BMI. The SLAM services use information comprised data on whether the person was an in-patient on 1 March 2020 and the number of days they were in contact with the SLAM services between 1 December 2019 and 1 March 2020. The contact with SLAM services included any form of in-patient and out-patient communication, such as email or phone or face-to-face consultations.

Statistical analysis

The data were analysed using STATA for Windows version 15.1. Using Cox proportional hazard models, we calculated hazard ratios for COVID-19-positive status in clozapine-treated participants compared with those treated with other antipsychotics. We censored observations at the date of death, date of COVID-19-positive test or 18 May 2020, whichever occurred first. We confirmed that the data satisfied the proportional-hazards assumptions using Schoenfeld residuals.

Three potential confounding variables contained missing data: smoking, ethnicity and BMI. First, we analysed the entire cohort using only variables with no missing data. Then, in a complete case analysis, we ran the same analyses, excluding individuals for

whom there were any missing data across any of the exposures investigated ($n = 5535$). Since the results were very similar, we were confident that the complete case analysis was unlikely to suffer from undue selection bias. We present results from the complete case analysis and the results of the whole-cohort analysis in the supplementary material available at https://doi.org/10.1192/bjp.2020.151.

Crude and adjusted models were constructed, first controlling for age, gender and ethnicity; then controlling for age, gender, ethnicity, in-patient status and number of contact days with the SLAM services. Last, we constructed a fully adjusted model controlling for all variables: age, gender, ethnicity, in-patient status, number of contact days with the SLAM services, smoking status and BMI. All models were built using data from the 5535 individuals with complete data. The above models were repeated using the whole cohort ($n = 6309$) without including the variables with missing data (ethnicity, smoking status, BMI): these results are in supplementary Table 1.

Results

There were 6309 active patients with schizophrenia-spectrum disorders (F2*) who were receiving any type of antipsychotic treatment during the beginning of the follow-up period. The sample mean age was 46.5 years (s.d. = 14.8) and men account for 61.7% of the sample. The sample's ethnic description is: 33.2% White (including White British, Irish or any other White background), 50.6% Black (including Black African, Black Caribbean, Other Black background, White and African, and White and Caribbean), 13.7% any Asian and Other ethnic background; 2.5% had missing data on ethnicity.

Table 1 summarises the demographic features of all the SLAM patients who qualified for the inclusion criteria ($n = 6309$). Of the individuals who were on clozapine, 66% were male, 46% were Black, 80% were current smokers and 48% had high BMI (obese). Compared with participants not on clozapine treatment, a higher proportion of clozapine-treated participants were in-patients in the hospital on 1 March 2020 (13 *v.* 6%), and clozapine-treated participants had more contact days with the SLAM services in the previous 3 months.

Table 2 summarises the demographic features presented according to their outcome status: COVID-19 positive or not COVID-19 positive. Of those who were COVID-19 positive, 41% were receiving clozapine treatment, whereas of those who were not COVID-19 positive, only 20% were receiving clozapine treatment. A higher proportion of COVID-19-positive patients were in-patients and COVID-19-positive patients had more contact days with the SLAM services.

The Cox regression analysis was performed with data of the 5535 individuals with complete information (774 participants were excluded because of missing data: Table 1), and the mean follow-up period was 78.00 days (s.d. = 7.03). Of these 5535 individuals, 92 tested positive for infection with COVID-19 during the follow-up period. Table 3 shows the hazard ratios for COVID-19 infection associated with being on clozapine-treatment in the

Table 1 Sample characteristics of all SLAM patients who qualified for the inclusion criteria, presented according to those who were and were not on clozapine treatment

	Not on clozapine treatment, *n* (%)	On clozapine treatment, *n* (%)
Total ($n = 6309$)	5027 (79.68)	1282 (20.32)
Gender		
Male	3047 (60.61)	847 (66.07)
Female	1980 (39.39)	435 (33.93)
Age		
<29 years	817 (16.25)	131 (10.22)
30–39 years	1055 (20.99)	281 (21.92)
40–49 years	1082 (21.52)	338 (26.37)
50–59 years	1145 (22.78)	369 (28.78)
60–69 years	546 (10.86)	127 (9.91)
≥70 years	382 (7.60)	36 (2.81)
Ethnicity		
White	1579 (31.41)	516 (40.25)
Black	2607 (51.86)	586 (45.71)
Asian, Other and Not stated	691 (13.75)	170 (13.26)
Missing	150 (2.98)	10 (0.78)
In-patient on the first day of follow-up period[a]		
No	4735 (94.19)	1113 (86.82)
Yes	292 (5.81)	169 (13.18)
SLAM service contact during index period[b]		
<4 days	1902 (37.84)	242 (18.88)
4–7 days	1575 (31.33)	479 (37.36)
≥8 days	1550 (30.83)	561 (43.76)
Smoking status		
Current smoker	3154 (62.74)	1028 (80.19)
Past smoker	1457 (28.98)	232 (18.10)
Never smoked	326 (6.48)	18 (1.40)
Missing	90 (1.79)	4 (0.31)
BMI		
Underweight and healthy	1601 (31.85)	264 (20.59)
Overweight	1335 (26.56)	358 (27.93)
Obese	1499 (29.82)	610 (47.58)
Missing	592 (11.78)	50 (3.90)

SLAM, South London and Maudsley NHS Foundation Trust; BMI, body mass index.
a. First date of follow period is 1 March 2020.
b. Index period is between 1 December 2019 and 1 March 2020, which is 3 months prior to the follow-up period.

Table 2 Sample characteristics of all SLAM patients who qualified for the inclusion criteria, presented according to those who tested positive for COVID-19 and those who did not during the follow-up period (1 March to 18 May 2020 inclusive)

	Not COVID-19 positive, *n* (%)	COVID-19 positive, *n* (%)
Total sample (*n* = 6309)	6207 (98.38)	102 (1.62)
On clozapine treatment		
No	4967 (80.02)	60 (58.82)
Yes	1240 (19.98)	42 (41.18)
Gender		
Male	3838 (61.83)	56 (54.90)
Female	2369 (38.17)	46 (45.10)
Age		
<29 years	937 (15.10)	11 (10.78)
30–39 years	1315 (21.19)	21 (20.59)
40–49 years	1408 (22.68)	12 (11.76)
50–59 years	1485 (23.92)	29 (28.43)
60–69 years	657 (10.58)	16 (15.69)
≥70 years	405 (6.52)	13 (12.75)
Ethnicity		
White	2069 (33.33)	26 (25.49)
Black	3129 (50.41)	64 (62.75)
Asian, Other and Not stated	853 (13.74)	8 (7.84)
Missing	156 (2.51)	4 (3.92)
In-patient on the first day of follow-up period[a]		
No	5804 (93.51)	44 (43.14)
Yes	403 (6.49)	58 (56.86)
SLAM service contact during index period[b]		
<4 days	2137 (34.43)	7 (6.86)
4–7 days	2035 (32.79)	19 (18.63)
≥8 days	2035 (32.79))	76 (74.51)
Smoking status		
Current smoker	4104 (66.12)	78 (76.47)
Past smoker	1669 (26.89)	20 (19.61)
Never smoked	342 (5.51)	2 (1.96)
Missing	92 (1.48)	2 (1.96)
BMI		
Underweight and healthy	1843 (29.69)	22 (21.57)
Overweight	1675 (26.99)	18 (17.65)
Obese	2054 (33.09)	55 (53.92)
Missing	635 (10.23)	7 (6.86)

SLAM, South London and Maudsley NHS Foundation Trust; BMI, body mass index.
a. First date of follow period is 1 March 2020.
b. Index period is between 1 December 2019 and 1 March 2020, which is 3 months prior to the follow-up period.

crude and adjusted models. The crude model shows a hazard ratio of 2.62 (95% CI 1.73–3.96) for participants receiving clozapine treatment and COVID-19 positive. This increased to 3.06 (95% CI 2.01–4.67) after adjusting for sociodemographic factors (age, gender, ethnicity). It was attenuated to 1.85 (95% CI 1.20–2.85) after adjusting for in-patient status and SLAM service contact. It was further attenuated to 1.76 (95% CI 1.14–2.72) after adjusting for BMI and smoking status.

Discussion

Summary of findings

Our findings suggest that receiving clozapine treatment is associated with increased risk of COVID-19 infection, compared with receiving any other type of antipsychotic treatment. Crude associations were attenuated but not completely explained by differences in sociodemographic factors such as age, gender and ethnicity, factors related to health conditions such as smoking status, BMI or proxies of availability of COVID testing (in-patient status or number of contacts with the SLAM services).

Comparison with previous studies

To our knowledge, no previous research has specifically investigated the associations between infection with COVID-19 and receiving clozapine treatment, as compared with receiving treatment with other antipsychotics.

In previous research, the risk of COVID-19 infection has been reported to be associated with older age, male gender, ethnicity (having an African, Caribbean, Other Black background, Bangladeshi or Pakistani background, or Indian (if male)) and higher BMI.[24,25] We found that older age was associated with COVID-19 infection and that infection rates were higher among Black people (compared with White people) and among people with high BMI (obese), but there were no significant associations with gender in our investigation.

Strengths

The cohort was large and inclusive of all patients who met the inclusion criteria in a defined population. SLAM is a near-monopoly provider for all aspects of secondary mental healthcare to a defined catchment area, so the study represents an almost comprehensive coverage of patients receiving clozapine treatment living in this catchment area of 1.3 million people.

In this analysis, the CRIS database made it possible to explore the complete electronic clinical records of more than 6000 individuals who met our inclusion criteria, which gave us the statistical power to be able to analyse a relatively rare event, and adjust for a range of potential confounders.

Table 3 Multivariate Cox analysis of association between receiving clozapine treatment and COVID-19 infection between 1 March and 18 May 2020 inclusive in 5535 individuals (92 COVID-19 positive)

Risk factors	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
On clozapine treatment				
No	1.00	1.00	1.00	1.00
Yes	2.62 (1.73–3.96)	3.06 (2.01–4.67)	1.85 (1.20–2.85)	1.76 (1.14–2.72)
Gender				
Male		1.00	1.00	1.00
Female		1.30 (0.86–1.97)	1.43 (0.94–2.17)	1.32 (0.86–2.04)
Age				
<29 years		1.00	1.00	1.00
30–39 years		1.03 (0.49–2.18)	1.25 (0.59–2.66)	1.14 (0.53–2.42)
40–49 years		0.54 (0.23–1.27)	0.69 (0.29–1.63)	0.64 (0.27–1.52)
50–59 years		1.30 (0.64–2.64)	1.95 (0.95–3.98)	1.78 (0.87–3.64)
60–69 years		1.73 (0.78–3.83)	2.88 (1.29–6.45)	2.76 (1.22–6.23)
≥70 years		3.31 (1.43–7.66)	4.10 (1.76–9.51)	4.35 (1.85–10.26)
Ethnicity				
White		1.00	1.00	1.00
Black		1.98 (1.23–3.19)	1.81 (1.12–2.91)	1.70 (1.04–2.77)
Asian, Other and Not stated		0.74 (0.30–1.82)	0.75 (0.31–1.84)	0.78 (0.32–1.92)
In-patient on the first day of follow-up period[a]				
No			1.00	1.00
Yes			10.31 (6.01–17.67)	10.08 (5.86–17.34)
SLAM service contact over index period[b]				
<4 days			1.00	1.00
4–7 days			2.58 (1.02–6.53)	2.62 (1.03–6.65)
≥8 days			3.56 (1.41–9.01)	3.64 (1.42–9.30)
Smoking status				
Current smoker				1.00
Past smoker				1.06 (0.62–1.81)
never smoked				0.69 (0.16–2.88)
BMI				
Underweight and healthy				1.00
Overweight				0.93 (0.49–1.75)
Obese				1.86 (1.11–3.14)

SLAM, South London and Maudsley NHS Foundation Trust; BMI, body mass index; HR, hazard ratio.
a. First date of follow period is 1 March 2020.
b. Index period is between 1 December 2019 and 1 March 2020, which is 3 months prior to the follow-up period.

In cohort studies, it is often impossible to be certain that the cases identified are true incident cases as opposed to prevalent cases that are identified during the study period. However, because there had been no cases of COVID-19 in SLAM at the start of the follow-up period, we can be certain that these are all incident cases of COVID-19. Furthermore, we can completely rule out reverse causation: the prescription of clozapine could not have been affected by knowledge of COVID-19 status since clozapine status was measured before any cases of COVID-19 had been diagnosed. Similarly, contact with services and in-patient status were measured before the start of the epidemic, so could not have been affected by COVID-19 status.

Limitations

We controlled for a number of potential confounders; however, there may still be residual confounding. There is a very large effect of in-patient status on the risk of COVID-19 infection. This is likely to arise partly from a higher risk of exposure to COVID-19 in hospital settings, and largely from the policy that in-patients showing any symptoms of COVID-19 were tested, while testing in the community was less comprehensive. Controlling for in-patient status on 1 March 2020 has not annulled the significant association between clozapine and COVID-19 infection. However, we cannot rule out the possibility that clozapine-treated patients could be more likely to be tested for COVID-19, even after accounting for the differences in patient contact and in-patient status between the groups before 1 March 2020. Also, it is possible that clozapine-treated patients might be more likely to be symptomatic with COVID-19, possibly owing to a reduced immune response, and therefore more likely to be tested. Consequently, a conservative interpretation of these findings might be that people on clozapine treatment are more likely to suffer from *symptomatic* COVID-19 infection, which is itself important clinically.

During the study period, SLAM enacted a policy of attempting to discharge patients back into the community where possible, to free up in-patient capacity. We are making an assumption that the proportion of patients discharged did not differ between the clozapine-treated group and the non-clozapine-treated group, the in-person and remote patient monitoring did not differ between the two groups, and that the amount of care and monitoring before compared with during the pandemic remained proportional between groups. Other potential confounders, such as cardiovascular diseases, hypertension, respiratory diseases or metabolic side-effects such as obesity and diabetes, were not included in the study because reliable data were not available for the whole cohort.

The most recent BMI measurements for some patients in the study were from almost 15 years ago. Although this is likely to give some indication of their BMI at the time of the study, it is important to note that BMI is more likely to have been recently measured in the clozapine-treated participants owing to the increased monitoring.

The 'current smoker' status of smoking data was extracted on the basis of the status within the 12 months prior to the follow-up period, so some of those data were from almost a year ago. Some of the 'past smoker' and 'never smoked' data were from almost 18 years ago. Given the impact of smoking on clozapine metabolisation and clozapine plasma levels, we cannot rule out that clozapine-

treated patients may be questioned more frequently about smoking and therefore have more up-to-date information regarding smoking habits.

Implications

To our knowledge, our results are the first to suggest that people on clozapine treatment are at higher risk of infection by COVID-19.[21] This is consistent with previous research demonstrating that people treated with clozapine have higher rates of infection and pneumonia than those on other antipsychotics and have alterations in both innate and adaptive immunity. There are also several alternative explanations for these findings, most notably the fact that clozapine-treated patients are likely to come into greater contact with services than patients on other antipsychotics and are therefore more likely to be tested if they develop symptoms. We have tried to adjust for patient contact, but, given the very large association between in-patient status and infection with COVID-19, we cannot confidently exclude the possibility that the association is explained by residual confounding.

The study is based on a relatively small number of cases, and we would not advocate any change in practice based on these findings alone. However, if the association is replicated and becomes firmly established, clinicians and patients will need to weigh up the increased risk of COVID-19 infection against the risk of psychotic relapse if clozapine is discontinued. Given that, for many patients, clozapine is the only effective antipsychotic, and with the well-established association between clozapine treatment and reduced all-cause mortality, these decisions are likely to be finely balanced and must be taken on a case-by-case basis.

Until this association is more firmly established, we would recommend that clinicians follow consensus guidelines for clozapine treatment during the COVID-19 pandemic, such as those of Siskind and colleagues[21] and Luykx and colleagues.[26] There should also be a focus on ensuring that clozapine-treated patients follow simple hygiene measures that can be taken to reduce the risks of COVID-19 infection, including handwashing, social distancing and the rigorous use of face masks and other personal protective equipment in clinical settings.

Future research

As the COVID-19 pandemic progresses, we and other groups will be able to study this association in larger samples and perhaps with better control of confounding. It will also be important to establish whether, among psychiatric patients with COVID-19, those treated with clozapine are at differential risk of adverse outcomes such as hospital admission, pneumonia, treatment in intensive care or ventilation, or death.

Risha Govind, MSc, Institute of Psychiatry, Psychology and Neuroscience, King's College London; and National Institute for Health Research (NIHR) Biomedical Research Centre, South London and Maudsley NHS Foundation Trust and King's College London, UK; **Daniela Fonseca de Freitas**, PhD, Institute of Psychiatry, Psychology and Neuroscience, King's College London; and National Institute for Health Research (NIHR) Biomedical Research Centre, South London and Maudsley NHS Foundation Trust and King's College London, UK; **Megan Pritchard**, MSc, Institute of Psychiatry, Psychology and Neuroscience, King's College London; and National Institute for Health Research (NIHR) Biomedical Research Centre, South London and Maudsley NHS Foundation Trust and King's College London, UK; **Richard D. Hayes**, PhD, Institute of Psychiatry, Psychology and Neuroscience, King's College London; and National Institute for Health Research (NIHR) Biomedical Research Centre, South London and Maudsley NHS Foundation Trust and King's College London, UK; **James H. MacCabe**, PhD, FRCPsych, Institute of Psychiatry, Psychology and Neuroscience, King's College London; and National Institute for Health Research (NIHR) Biomedical Research Centre, South London and Maudsley NHS Foundation Trust and King's College London; and National Psychosis Unit, South London and Maudsley NHS Foundation Trust, London, UK

Correspondence: James MacCabe. Email: james.maccabe@kcl.ac.uk

First received 22 May 2020, final revision 10 Jul 2020, accepted 21 Jul 2020

Supplementary material

Supplementary material is available online at http://doi.org/10.1192/bjp.2020.151.

Data availability

The study used clinical data held by South London and Maudsley NHS Foundation Trust. The data are not available to those outside this organisation.

Author contributions

J.H.M. formulated the research question. All authors were involved in data analysis and in conducting the study. All authors contributed to the manuscript and approved the submitted version. R.D.H. and J.H.M. take academic responsibility for the work.

Funding

This work was supported by the Clinical Records Interactive Search (CRIS) system, funded and developed by the National Institute for Health Research (NIHR) Mental Health Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, and a joint infrastructure grant from Guy's and St Thomas' Charity and the Maudsley Charity (grant number BRC-2011-10035). All authors receive salary support from the NIHR Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR or the Department of Health. The funders had no role in the study design; in the collection, analysis and interpretation of the data; in the writing of the report; and in the decision to submit the paper for publication.

Declaration of interest

R.D.H. has received research funding from Roche, Pfizer, Janssen and Lundbeck. D.F.F. has received research funding from Janssen and Lundbeck. J.H.M. has received research funding from Lundbeck.

ICMJE forms are in the supplementary material, available online at https://doi.org/10.1192/bjp.2020.151.

References

1 Siskind D, McCartney L, Goldschlager R, Kisely S. Clozapine *v*. first- and second-generation antipsychotics in treatment-refractory schizophrenia: systematic review and meta-analysis. *Br J Psychiatry* 2016; **209**: 385–92.

2 Cho J, Hayes RD, Jewell A, Kadra G, Shetty H, MacCabe JH, et al. Clozapine and all-cause mortality in treatment-resistant schizophrenia: a historical cohort study. *Acta Psychiatr Scand* 2018; **139**: 237–47.

3 Kesserwani J, Kadra G, Downs J, Shetty H, MacCabe JH, Taylor D, et al. Risk of readmission in patients with schizophrenia and schizoaffective disorder newly prescribed clozapine. *J Psychopharmacol* 2019; **33**: 449–58.

4 Wimberley T, MacCabe JH, Laursen TM, Sørensen HJ, Astrup A, Horsdal HT, et al. Mortality and self-harm in association with clozapine in treatment-resistant schizophrenia. *Am J Psychiatry* 2017; **174**: 990–8.

5 Hayes RD, Downs J, Chang CK, Jackson RG, Shetty H, Broadbent M, et al. The effect of clozapine on premature mortality: an assessment of clinical monitoring and other potential confounders. *Schizophr Bull* 2015; **41**: 644–55.

6 Hayes JF, Marston L, Walters K, King MB, Osborn DPJ. Mortality gap for people with bipolar disorder and schizophrenia: UK-based cohort study 2000–2014. *Br J Psychiatry* 2017; **211**: 175–81.

7 Vermeulen JM, Van Rooijen G, Van De Kerkhof MPJ, Sutterland AL, Correll CU, De Haan L. Clozapine and long-term mortality risk in patients with schizophrenia: a systematic review and meta-analysis of studies lasting 1.1–12.5 years. *Schizophr Bull* 2019; **45**: 315–29.

8 Chou FHC, Tsai KY, Chou YM. The incidence and all-cause mortality of pneumonia in patients with schizophrenia: a nine-year follow-up study. *J Psychiatr Res* 2013; **47**: 460–6.

9 Seminog OO, Goldacre MJ. Risk of pneumonia and pneumococcal disease in people with severe mental illness: English record linkage studies. *Thorax* 2013; **68**: 171–6.

10 Shen TC, Chen CH, Huang YJ, Lin CL, Chang TC, Tu CY, et al. Risk of pleural empyema in patients with schizophrenia: a nationwide propensity-matched cohort study in Taiwan. *BMJ Open* 2018; **8**(7): e021187.

11 John A, McGregor J, Jones I, Lee SC, Walters JTR, Owen MJ, et al. Premature mortality among people with severe mental illness: new evidence from linked primary care data. *Schizophr Res* 2018; **199**: 154–62.

12 Gurillo P, Jauhar S, Murray RM, MacCabe JH. Does tobacco use cause psychosis? Systematic review and meta-analysis. *Lancet Psychiatry* 2015; **2**: 718–25.

13 Haddad PM. Current use of second-generation antipsychotics may increase risk of pneumonia in people with schizophrenia. *Evid Based Ment Health* 2013; **16**(4): 109.

14 Kuo C-J, Yang S-Y, Liao Y-T, Chen WJ, Lee W-C, Shau W-Y, et al. Second-generation antipsychotic medications and risk of pneumonia in schizophrenia. *Schizophr Bull* 2013; **39**: 648–57.

15 Stoecker ZR, George WT, O'Brien JB, Jancik J, Colon E, Rasimas JJ. Clozapine usage increases the incidence of pneumonia compared with risperidone and the general population: a retrospective comparison of clozapine, risperidone, and the general population in a single hospital over 25 months. *Int Clin Psychopharmacol* 2017; **32**: 155–60.

16 De Leon J, Sanz EJ, De las Cuevas C. Data from the World Health Organization's pharmacovigilance database supports the prominent role of pneumonia in mortality associated with clozapine adverse drug reactions. *Schizophr Bull* 2020; **46**: 1–3.

17 Liu NH, Daumit GL, Dua T, Aquila R, Charlson F, Cuijpers P, et al. Excess mortality in persons with severe mental disorders: a multilevel intervention framework and priorities for clinical practice, policy and research agendas. *World Psychiatry* 2017; **16**: 30–40.

18 Gurrera RJ, Perry NL. Clozapine-associated aspiration pneumonia: case series and review of the literature: reply. *Psychosomatics* 2019; **60**: 103.

19 Li XH, Zhong XM, Lu L, Zheng W, Wang SB, Rao WW, et al. The prevalence of agranulocytosis and related death in clozapine-treated patients: a comprehensive meta-analysis of observational studies. *Psychol Med* 2020; **50**: 583–94.

20 Ponsford M, Castle D, Tahir T, Robinson R, Wade W, Steven R, et al. Clozapine is associated with secondary antibody deficiency. *Br J Psychiatry* 2019; **214**: 83–9.

21 Siskind D, Honer WG, Clark S, Correll CU, Hasan A, Howes O, et al. Consensus statement on the use of clozapine during the COVID-19 pandemic. *J Psychiatry Neurosci* 2020; **45**: 222–3.

22 Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, et al. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry* 2009; **9**: 51.

23 CRIS NLP Applications Library. *CRIS Natural Language Processing v1.1*. National Institute for Health Research Biomedical Research Centre, 2020 (https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/cris-natural-language-processing/).

24 Liu K, Chen Y, Lin R, Han K. Clinical features of COVID-19 in elderly patients: a comparison with young and middle-aged patients. *J Infect* 2020; **80**(6): e14–8.

25 White C, Nafilyan V. *Coronavirus (COVID-19) Related Deaths by Ethnic Group, England and Wales: 2 March 2020 to 10 April 2020*. Office for National Statistics, 2020 (https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/coronavirusrelateddeathsbyethnicgroupenglandandwales/2march2020to10april2020).

26 Luykx JJ, van Veen SMP, Risselada A, Naarding P, Tijdink JK, Vinkers C. Safe and informed prescribing of psychotropic medication during the COVID-19 pandemic. *Br J Psychiatry* [Epub ahead of print] 4 May 2020. Available from: https://doi.org/10.1192/bjp.2020.92.

CHAPTER 6

6 PUBLICATION: COVID-RELATED HOSPITALIZATION, INTENSIVE CARE TREATMENT, AND ALL-CAUSE MORTALITY IN PATIENTS WITH PSYCHOSIS AND TREATED WITH CLOZAPINE

This chapter incorporates the following publication:

> Govind, R. et al. (2022) 'COVID-related hospitalization, intensive care treatment, and all-cause mortality in patients with psychosis and treated with clozapine', *European Neuropsychopharmacology.* Elsevier. doi: 10.1016/J.EURONEURO.2022.01.007.

Contribution Statement

Daniela Fonseca de Freitas, Richard Hayes, James MacCabe and I conceived and designed this study. With the support and guidance from other authors, I performed all the data extraction and statistical analysis for this chapter; Megan Pritchard supervised the data extraction part; Daniela Fonseca de Freitas and Mizanur Khondoker supervised the statistical analysis part. I also wrote the manuscript. All authors critically reviewed the manuscript and approving the final version.

European Neuropsychopharmacology 56 (2022) 92-99

www.elsevier.com/locate/euroneuro

COVID-related hospitalization, intensive care treatment, and all-cause mortality in patients with psychosis and treated with clozapine

Risha Govind[a,b], Daniela Fonseca de Freitas[a,b], Megan Pritchard[a,b], Mizanur Khondoker[c], James T. Teo[a,d], Robert Stewart[a,b], Richard D. Hayes[a,b,#], James H. MacCabe[a,b,e,#,*]

[a] *Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom*
[b] *South London and Maudsley NHS Foundation Trust, London, United Kingdom*
[c] *Faculty of Medicine and Health Sciences, Norwich Medical School, University of East Anglia, Norwich, United Kingdom*
[d] *King's College Hospital, Denmark Hill, London, United Kingdom*
[e] *National Psychosis Unit, South London and Maudsley NHS Foundation Trust, London, United Kingdom*

Received 15 October 2021; received in revised form 14 January 2022; accepted 18 January 2022

KEYWORDS
Antipsychotic agents;
Clozapine;
COVID-19;
SARS-CoV-2;
Schizophrenia

Abstract
Clozapine, an antipsychotic, is associated with increased susceptibility to infection with COVID-19, compared to other antipsychotics. Here, we investigate associations between clozapine treatment and increased risk of adverse outcomes of COVID-19, namely COVID-related hospitalisation, intensive care treatment, and death, amongst patients taking antipsychotics with schizophrenia-spectrum disorders. Using the clinical records of South London and Maudsley NHS Foundation Trust, we identified 157 individuals who had an ICD-10 diagnosis of schizophrenia-spectrum disorders, were taking antipsychotics (clozapine or other antipsychotics) at the time of COVID-19 pandemic in the UK and had a laboratory-confirmed COVID-19 infection. The following health outcomes were measured: COVID-related hospitalisation, COVID-related intensive care treatment and death. We tested associations between clozapine treatment and

* Corresponding author at: King's College London Institute of Psychiatry Psychology and Neuroscience, United Kingdom.
E-mail address: james.maccabe@kcl.ac.uk (J.H. MacCabe).
[#] Authors contributed equally

https://doi.org/10.1016/j.euroneuro.2022.01.007
0924-977X/

each outcome using logistic regression models, adjusting for gender, age, ethnicity, neighbourhood deprivation, obesity, smoking status, diabetes, asthma, bronchitis and hypertension using propensity scores. Of the 157 individuals who developed COVID-19 while on antipsychotics (clozapine or other antipsychotics), there were 28% COVID-related hospitalisations, 8% COVID-related intensive care treatments and 8% deaths of any cause during the 28 days follow-up period. amongst those taking clozapine, there were 25% COVID-related hospitalisations, 7% COVID-related intensive care treatments and 7% deaths. In both unadjusted and adjusted analyses, we found no significant association between clozapine and any of the outcomes. Thus, we found no evidence that patients with clozapine treatment at time of COVID-19 infection had increased risk of hospitalisation, intensive care treatment or death, compared to non-clozapine antipsychotic-treated patients. However, further research should be considered in larger samples to confirm this.

1. Introduction

Clozapine is an atypical antipsychotic, the gold standard drug for treatment-resistant schizophrenia, and the only effective treatment for many patients with schizophrenia (Siskind et al., 2016). Patients with schizophrenia have a higher risk for developing pneumonia and, compared to the general population, have higher premature mortality (Chou, Tsai and Chou, 2013; Seminog and Goldacre, 2013; Hayes et al., 2017; John et al., 2018; Shen et al., 2018; Vermeulen et al., 2019). Patients receiving clozapine treatment have lower rates of overall hospitalisation and mortality compared to those receiving other antipsychotic treatments (Hayes et al., 2015; Wimberley et al., 2017; Cho et al., 2018; Kesserwani et al., 2019). However, clozapine is associated with an increased risk of developing pneumonia (Haddad, 2013; Kuo et al., 2013; Stoecker et al., 2017; De Leon, Sanz and De las Cuevas, 2020). This might be explained by confounding by indication, in that clozapine is predominantly prescribed in cases of treatment-resistant schizophrenia, associated in itself with higher rates of comorbidities such as smoking cigarettes, inadequate physical activity, and poor diet (Liu et al., 2017). Alternatively, clozapine could increase the risk of pneumonia via immunosuppression, or via other adverse effects of clozapine which could fall on the causal pathway, such as hypersalivation (causing aspiration pneumonia), diabetes and obesity (Newcomer, 2005; Liu et al., 2017; De Leon, Sanz and De las Cuevas, 2020).

COVID-19 first appeared in China in December 2019 and was declared a global pandemic by the WHO in March 2020 (Siskind et al., 2020). It is caused by the SARS-Cov2 virus, and has pathological effects on multiple organ systems including the lungs, heart, brain, kidney, gastrointestinal tract, liver and spleen (Tabary et al., 2020). The most concerning consequence of the infection is respiratory failure. The most severe cases of COVID-19 can require hospitalisation and treatment in intensive care, and mortality is significant. Several studies have been performed to investigate the impact of COVID-19 on patients on clozapine treatment (Govind et al., 2020; Vita and Barlati, 2021). In a previous study, we reported that patients on clozapine treatment may be at higher risk of COVID-19 infection (Govind et al., 2020). Recently, case studies on this have been presented by Butler et al., Boland and Dratcu (Boland and Dratcu, 2020; Butler et al., 2020); however, to our knowledge, the association between clozapine treatment and the adverse outcomes of COVID-19 have yet to be investigated in an epidemiological sample.

In this paper, we investigated whether clozapine treatment, compared to non-clozapine antipsychotic treatment, at time of COVID-19 infection, was associated with an increased risk of adverse outcomes of COVID-19 in patients with schizophrenia in a geographically defined population in London during the COVID-19 pandemic in the UK.

2. Method

2.1. Setting and ethics statement

A retrospective cohort study was carried out using data from the electronic records of the South London and Maudsley NHS Foundation Trust (SLAM). SLAM caters to all secondary mental health care needs of over 1.3 million people of four London boroughs (Lambeth, Southwark, Lewisham, and Croydon). SLAM has used a fully electronic clinical records system since 2006, and the Clinical Records Interactive Search (CRIS) platform was established to render full, de-identified clinical records available to researchers for secondary analysis within robust data security and governance framework (Stewart et al., 2009). CRIS was approved for use as a de-identified data resource for secondary analysis by Oxfordshire Research Ethics Committee C (reference 18/SC/0372).

CRIS includes both structured and free-text fields from the clinical notes, and custom-built Natural Language Processing (NLP) algorithms are used to extract information from the latter, the specifications and performance metrics of which are detailed in an open online catalogue (CRIS NLP Applications Library, 2020). Data from four NLP algorithms were used in this study: diagnosis, medication, smoking and body mass index (BMI). Information regarding COVID-19 patient cases admitted to two local hospitals (King's College Hospital and Princess Royal University Hospital) were obtained via a data linkage (performed under Regulation 3(2) and Regulation 3(3) of the Health Service Control of Patient Information Regulations 2002 (COPI)).

2.2. Cohort

The cohort comprised individuals who satisfied all three of the following inclusion criteria: (1) a laboratory-confirmed COVID-19 infection between March 01, 2020, and December 20, 2020; (2)

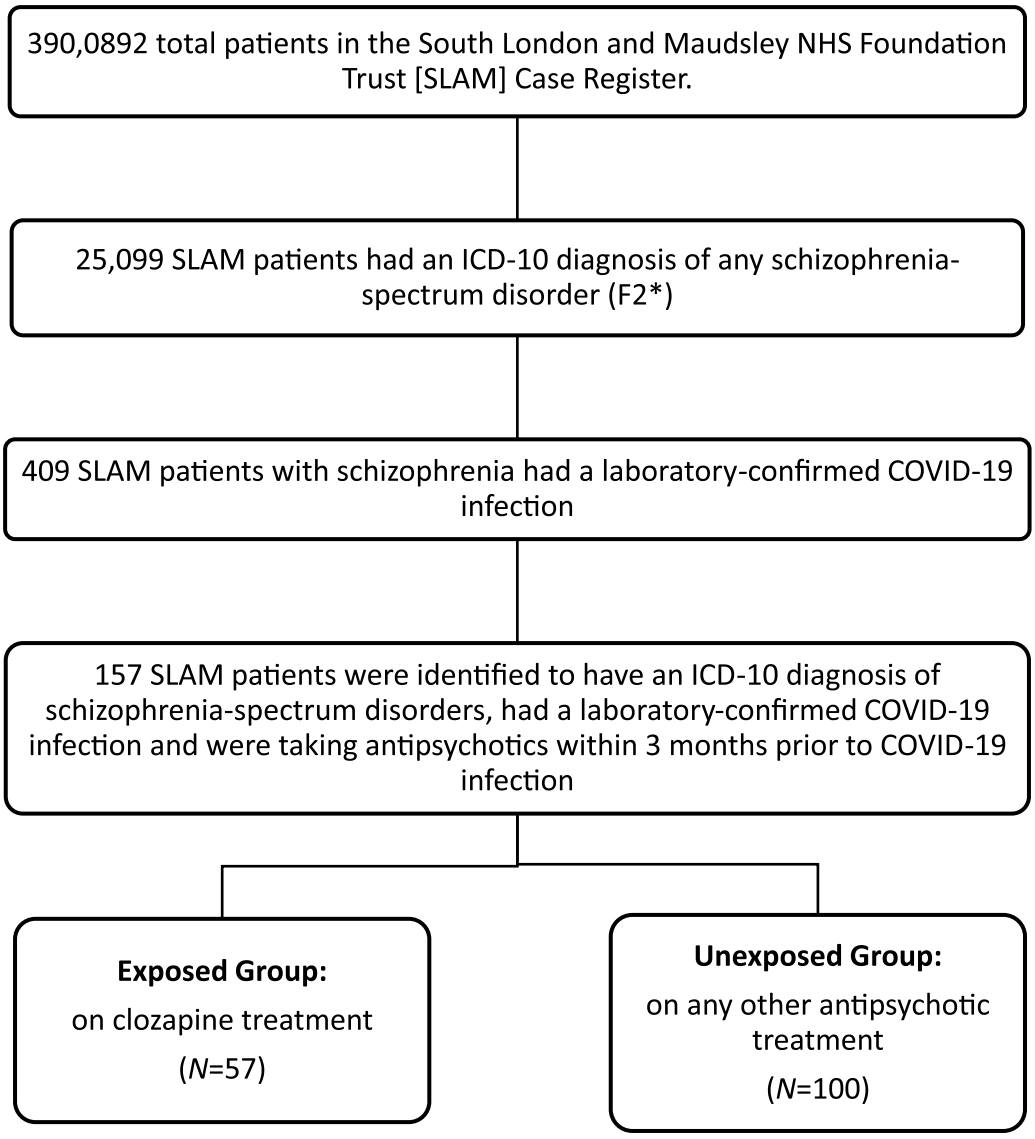


Fig. 1 Study design.

ICD-10 diagnosis of any schizophrenia-spectrum disorder (F2*); (3) recorded as taking antipsychotic medication within 3 months prior to the date of COVID-19 infection. Figure 1 shows the study design. SQL Server Management Studio version 15.0 (Microsoft Inc, USA) was used to extract the data. The day of data extraction was January 07, 2021. Patients were followed-up from the date of COVID-19 infection until they were hospitalised, entered intensive care treatment, died, or reached the end of the follow-up period (within 28 days of infection). Since we did not have access to the cause of death information, measured all cause mortality within 28 days of COVID-19 diagnosis, in line with Public Health England (Department of Health and Social Care, 2020).

Diagnosis of schizophrenia-spectrum disorder (ICD-10: F2*) was ascertained via a diagnosis algorithm, by which NLP outputs are combined with data in the structured fields, such as the data from ICD-10 diagnosis forms in the source record (CRIS NLP Applications Library, 2020).

Antipsychotic medication within 3 months prior to COVID-19 infection was identified by an NLP algorithm that targeted administrations of 29 different antipsychotic medications (Perera et al., 2016; CRIS NLP Applications Library, 2020). The medications algorithm NLP outputs are combined with data from structured fields, including SLAM pharmacy dispensing data.

The COVID-19 infection data used for the inclusion criteria were collated by combining information from three sources: (1) SLAM pathology lab results data, (2) the presence of a clinician-entered alert on SLAM records indicating a positive test, and (3) data provided by local general hospitals (King's College Hospital and Princess Royal University Hospital) for COVID-19 related admissions. The COVID-19 infection dates were verified and, when needed, were rectified to the earliest mention of COVID-19-compatible symptoms or COVID-19 tests, according to the information presented in SLAM's clinical notes. To cater to scenarios where the COVID-19 test was conducted after an admission for symptomatic COVID, the COVID-19 infection date was changed to the date of hospital admission when the positive test result was within 7 days of hospital admission. The patients were removed from the analysis either if the clinical notes stated that their COVID-19 positive status was entered by mistake or they had COVID-19 infection after December 20, 2020.

2.3. Exposure of interest

People who were recorded as receiving clozapine treatment at any time within 3 months prior to the assigned COVID-19 infection date were defined as the exposed group. Those on any type or combination of antipsychotic treatment that did not include clozapine during this time constituted the unexposed group.

2.4. Main outcome measures

The outcomes of interest were: (1) COVID-related hospitalisation (2) COVID-related intensive care treatment, and (3) all-cause mortality during the follow-up period (within 28 days of COVID-19 infection). These data were collated by combining COVID-19 related information provided by local general hospitals (King's College Hospital and Princess Royal University Hospital) and the data in the SLAM records. The SLAM records data on hospitalisation and intensive care treatment were curated by reading the clinical notes of each patient from the date of COVID-19 infection until a positive mention of hospitalisation or mention of recovery from COVID-19. The SLAM records data on mortality were retrieved from structured fields in SLAM health records which are populated on weekly basis via linkage with the NHS Spine.

2.5. Potential confounding variables

We considered as potential confounding variables sociodemographic characteristics and behavioural/clinical factors. The sociodemographic information comprised age, gender, ethnicity, and neighbourhood deprivation. The behavioural/clinical factors were smoking status, obesity, diabetes, asthma, bronchitis and hypertension.

Age was calculated at the time of COVID-19 infection from the year and month of birth. Data on gender and ethnicity came from the routinely collected data in structured fields in SLAM health records. SLAM records include ethnicity in 14 categories, which were collapsed into 3 categories, "White", "Black" and "Asian & other". The category "White" was a conflation of White British, White Irish and White Other. The category "Black" was a conflation of Black African, Black Other (which comprises Black British), Black Caribbean, Mixed Race White and Black Caribbean and Mixed Race White and Black African. The category "Asian & Other" was a conflation of Indian, Pakistani, Other Asian, and Other ethnic group. For patients with no ethnicity information, including those with ethnicity as "not stated" in the structured fields, their ethnicity data was extracted by manually reviewing the record text fields.

Neighbourhood deprivation was measured using the Index of Multiple Deprivation (IMD) 2019 applying Census-derived data to the Lower Super Output Area: a standard administrative unit containing an average of 1500 residents. The deciles of the IMD range between 1, the most deprived, and 10, the least deprived. The data from IMD deciles 1 to 3 were merged to form the "Higher level of deprivation" category. The data from IMD deciles 4 to 10 were merged to form the "Lower level of deprivation" category. A third category, "homeless", was created for the patients who were homeless.

Smoking behaviour in the year prior to COVID-19 infection was identified using an NLP algorithm (CRIS NLP Applications Library, 2020), supplemented by a manual review of record text fields. Similarly, the obesity status was derived from recorded body mass index (BMI) scores ascertained via an NLP algorithm, supplemented by manual records text review, choosing the most recent extracted score prior to the COVID-19 infection date (CRIS NLP Applications Library, 2020). Obesity was defined as BMI is greater than or equal to 30 (World Health Organization, 1995). Data on physical illnesses (diabetes, asthma, bronchitis and hypertension) were extracted manually from relevant free-text fields of the patient records for each patient, aided by search strings.

2.6. Statistical analysis

The data were analysed using STATA for Windows version 15.1. Since the data on the date of COVID-19 infection which is the time zero date was not precise, we used logistic regression instead of Cox proportional hazard models for the analysis. In the unadjusted analysis, we used logistic regression to calculate odds ratios comparing clozapine treated patients to those treated with other antipsychotics for each of the outcomes described above. Covariate adjustment was made via propensity scores within a logistic regression model as direct adjustment for all covariates was not feasible due to limited sample size. The propensity scores were predicted from a separate logistic regression model using clozapine treatment as the outcome and the sociodemographic (age, gender, ethnicity, neighbourhood deprivation), behavioural/clinical factors (smoking status, BMI, diabetes, asthma, bronchitis, hypertension) as predictor variables. The logit (log-odds) of the probability of clozapine treatment (propensity score) was included as a single covariate along with the exposure (indicator of clozapine treatment) in the logistic regression models for adjusted analysis. STATA was also used to estimate power for the analysis.

3. Results

There were 157 patients ascertained with a laboratory-confirmed COVID-19 infection and schizophrenia-spectrum disorders (F2*) who were receiving any type of antipsychotic treatment during the study period. The follow-up period was 28 days after COVID-19 infection. The study sample comprised of patients treated with these antipsychotic medications: clozapine (36%), olanzapine (50%), risperidone (31%), aripiprazole (35%), amisulpride (12%), paliperidone (18%), flupentixol (13%), haloperidol (17%), zuclopenthixol (11%), quetiapine (10%), fluphenazine (3%), piportil (3%), sulpiride (3%), lurasidone (3%), trifluoperazine (1%), chlorpromazine (1%), pipotiazine (3%), penfluridol (1%) and droperidol (1%). The percentages refer to the proportion of the cohort on each antipsychotic and not all patients were on monotherapy. The mean age of the study sample was 50.6 years (SD=16.01), and men accounted for 53.5% of the sample. The study sample had a relatively high proportion of patients from minority ethnic groups: 61.2% Black, 10.8% any Asian and Other ethnic background and 28.0% White. Table 1 summarises the demographic features of all the SLAM patients who were eligible for inclusion based on the inclusion criteria (N=157). Of the individuals who were receiving clozapine, 56% were male, 58% were Black, 88% were current smokers, and 63% were obese.

Of the 157 individuals, 44 (28%) had an episode of COVID-related hospitalisation, 13 (8%) received COVID-related intensive care treatment and 13 (8%) died of any cause during the 28 days follow-up period. The majority of deaths were in people not admitted to intensive care: only 23% of those reporting being in intensive care, according to SLaM notes and data linkage to the two hospitals, died. amongst those taking clozapine, 25% had COVID-related hospitalisations, 7% had COVID-related intensive care treatments and 7% died. amongst those not taking clozapine, there were

Table 1 Sample description of the 157 SLAM patients who qualified for the inclusion criteria, presented according to those who were and were not receiving clozapine-treatment.

	On Clozapine treatment (%)	Not on Clozapine-treatment (%)
Total Patients in cohort	36.3 (n=57)	63.7 (n=100)
Males	56.1	52.0.0
Age		
< 50 years	52.6	40.0
> 50 years	47.4	60.0
Ethnicity		
White	28.1	28.0
Black	57.9	63.0
Asian & Other	14.0	9.0
Neighbourhood Deprivation		
Lower level of deprivation	56.1	45.0
Higher level of deprivation	35.1	49.0
Homeless	8.8	6.0
Current smoker	87.7	58.0
Obesity	63.2	35.0
Diabetes	43.9	42.0
Asthma	28.1	16.0
Bronchitis	8.8	13.0
Hypertension	45.6	45.0
Outcomes		
COVID-19 hospitalisation	24.6	30.0
COVID-19 treatment in	7.0	9.0
intensive care	7.0	9.0
All-cause mortality		

Table 2 Logistic regression analysis of the association between receiving clozapine treatment and each outcome (COVID-related hospitalisation, COVID-related treatment in intensive care and death) between the date of COVID-19 infection and January 07, 2021, inclusive in 157 individuals.[a]

clozapine treatment risk factor for outcome:	COVID-19 hospitalisation Odds Ratio (95% CI)	COVID-19 treatment in intensive care Odds Ratio (95% CI)	All-cause mortality Odds Ratio (95% CI)
Unadjusted	0.76 (0.36-1.59)	0.76 (0.22-2.60)[b]	0.76 (0.22-2.60)[b]
Adjusted for confounding effects using propensity scores[c]	1.12 (0.48-2.60)	0.71 (0.18-2.77)	1.38 (0.33-5.71)

[a] Of the 157 individuals, 44 patients had COVID-related hospitalisation, 13 patients had COVID-related treatment in intensive care, 13 patients died after COVID-19 infection

[b] Since the number of patients who had COVID-related treatment in intensive were the same as the number of patients who died after COVID-19 infection, the unadjusted analysis for these two outcomes produced the same results.

[c] list of confounding variables: gender, age, ethnicity, neighbourhood deprivation, smoking status, obesity, diabetes, asthma, bronchitis and hypertension

30% COVID-related hospitalisations, 9% COVID-related intensive care treatments and 9% deaths.

The logistic regression analysis was performed on the 157 individuals for each of the three outcomes, and Table 2 shows the odds ratio for each in the unadjusted and propensity score adjusted models. In unadjusted analyses, receiving clozapine treatment was not significantly associated with any outcome. Furthermore, no significant association was observed for any of the outcomes after covariate adjustment. The unadjusted odds ratio for COVID-related hospitalisation was 0.76 (95% CI: 0.36-1.59), COVID-related treatment in intensive care was 0.76 (95% CI: 0.22-2.60) and all-cause mortality was 0.76 (95% CI: 0.22-2.60). Since the number of patients who had COVID-related treatment in intensive was the same as the number of patients who died after COVID-19 infection, the unadjusted analysis for these two outcomes produced the same results. The adjusted odds ratio for COVID-related hospitalisation was 1.12 (95% CI: 0.48-2.60), COVID-related treatment in intensive care was 0.71 (95% CI: 0.18-2.77) and all-cause mortality was 1.38 (95% CI: 0.33-5.71). Post-hoc power calculations indicated that the sample size was sufficient to detect with 80% power (alpha 0.05) an odds ratio of 2.78 for COVID-related hospitalisation and 3.95 for all-cause mortality. Given the known strong association between obesity and the risk of adverse outcomes in COVID-19 infection, we ran another model where we directly included obesity as a covariate in the propensity score adjusted model. We did not see

any material change in the results indicating propensity score adjustment has done a good job in accounting for the imbalance of obesity between the groups.

4. Discussion

4.1. Summary of findings

We investigated if receiving clozapine treatment, compared to non-clozapine antipsychotic treatment, may be associated with increased risk of hospitalisation, intensive care treatment or all-cause mortality (within 28 days from infection) in COVID-19 positive patients with schizophrenia-spectrum disorders. We found no evidence that receiving clozapine treatment substantially increases the risk of these outcomes, compared to receiving any other types of antipsychotic treatment.

4.2. Comparison with previous studies

To our knowledge, no previous research has specifically investigated the associations between receiving clozapine treatment, as compared to receiving treatment with other antipsychotics, and hospitalisation, intensive care treatment or mortality from COVID-19.

4.3. Strengths and limitations

As a strength of this study, SLAM is a near-monopoly service provider of all aspects of secondary mental health care to residents within a defined geographic catchment, allowing relatively comprehensive ascertainment of people with the disorders of interest receiving specialist care during the COVID-19 pandemic in the UK. The CRIS database provided the platform to ascertain the relevant sample and access information on a range of potential confounders. However, it is important to bear in mind that not all people with schizophrenia-spectrum disorders will have been receiving specialist mental healthcare at that time, so that generalisability is limited. The CRIS database does not include all antipsychotic medication data on patients who are discharged to the GP services and therefore even though these patients qualified for the inclusion criteria, they would have been missed. Also, some patients would have been missed because the cohort was extracted using results from NLP algorithms; the precision and recall scores for the diagnosis algorithm was 100% and 65% respectively; The precision and recall scores for the antipsychotics part of the medication algorithm was 88% and 90%, respectively (CRIS NLP Applications Library, 2020). Furthermore, not all COVID-19 infection episodes will have been ascertained, particularly during the early stages of the pandemic when access to tests was very limited.

The COVID-related hospitalisation and intensive care treatment data came from combining information provided by local general hospitals and supplementing that by reading the anonymised clinical notes of each patient. For the COVID-19 infections that were diagnosed in the catchment area, given that lockdown restrictions precluded travel away from one's primary residence, it is unlikely the patients would have travelled to other areas during this period and therefore, their diagnosis should have been recorded at local hospitals. Our data linkage included King's College Hospital and Princess Royal University Hospital, which are two large healthcare providers in the area. To cover diagnoses that were made in hospitals not included in our linkage, we included data from clinician-entered alerts on SLAM records. There maybe be some bias attributed to clozapine-treated patients being in more contact with the SLAM services (Govind et al., 2020). This would have resulted in complications from COVID-19 infections being more likely to be recorded in patients taking clozapine than in those taking other antipsychotics, biasing the results towards unfavourable outcomes in clozapine patients.

Another important limitation of the analysis results from type II error. We acknowledge this analysis is underpowered, but since the sample included all eligible patients so could not be increased any further, there was no rationale to conduct a power calculation prior to the study. We calculated the post-hoc power calculation to give the context in terms of the probability of a Type II error and to show that a large association was unlikely.

Although all deaths occurred within 28 days of COVID-19 infection, and thus meet the Public Heath England criteria for COVID-related deaths (Public Health England, 2020), it is possible that some deaths may have been unrelated to COVID.

Obesity is a recognised risk factor for more severe COVID-19 outcomes, but obesity information had to be extrapolated using the nearest BMI score, not all of which were recent. While these BMI scores are likely to give some indication of obesity of the patient, we cannot rule out that clozapine-treated patients may have more up-to-date BMI scores due to increased monitoring. Compared to those receiving treatment with other antipsychotics, a significantly high proportion of clozapine-treated patients were obese. This may be attributable to clozapine having the highest potential to induce weight gain compared to other antipsychotics (Allison et al., 1999).

Since the smoking data encompasses patients who have mentions of cigarette smoking in their clinical records from any time within a year prior to COVID-19 infection, some patients' smoking status may have been misclassified. Given the impact of smoking on clozapine metabolization and clozapine plasma levels, it is important to note that clozapine-treated patients are more likely to be questioned about their smoking habits and therefore have recent information on it.

4.4. Implications

To our knowledge, this is the first study to investigate whether clozapine-treated patients are at increased risk of adverse outcomes of COVID-19, such as hospitalisation, treatment in intensive care or ventilation, or all-cause mortality, than patients on other antipsychotics. Within the limits of statistical power, we did not find evidence of substantial increased risk; however, larger and/or multi-site studies would be needed to rule out smaller effects.

5. Author contributions

Conceived and design of the study: RG, DFF, RDH, JHM
Performed the analysis: RG
Interpreted the results, paper writing & critical review: all authors (RG, DFF, MP, MK, JTT, RS, RDH, JHM)
Gave final approved of the version to be published: all authors (RG, DFF, MP, MK, JTT, RS, RDH, JHM)

6. Role of funding

All authors receive salary support from the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RS is additionally part-funded by: i) a Medical Research Council (MRC) Mental Health Data Pathfinder Award to King's College London; ii) an NIHR Senior Investigator Award; iii) the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health. The above funding had no role in the study design; in the collection, analysis, and interpretation of the data; in the writing of the report; and in the decision to submit the paper for publication.

Declaration of Competing Interest

RDH has received research funding from Roche, Pfizer, Janssen, and Lundbeck. DFF has received research funding from Janssen and Lundbeck. JHM has received research funding from Lundbeck. JTT has received research funding from Bristol-Meyers-Squibb. RS declares research support in the last 36 months from Janssen, GSK and Takeda.

Acknowledgement

This work was supported by the Clinical Records Interactive Search (CRIS) system funded and developed by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London and a joint infrastructure grant from Guy's and St Thomas' Charity and the Maudsley Charity (grant number BRC-2011-10035).

Ethics statement

The research was conducted under ethical approval reference 18/SC/0372 from Oxfordshire Research Ethics Committee C.

References

Allison, D.B., et al., 1999. Antipsychotic-induced weight gain: A comprehensive research synthesis. Am. J. Psychiatry 156 (11), 1686–1696. doi:10.1176/ajp.156.11.1686.

Boland, X., Dratcu, L., 2020. 'Clozapine in the time of COVID-19', Clinical Psychopharmacology and Neuroscience. Korean College of Neuropsychopharmacology 18 (3), 450–453. doi:10.9758/CPN.2020.18.3.450.

Butler, M., et al., 2020. Clozapine prescribing in COVID-19 positive medical inpatients: a case series. Therapeutic Advances in Psychopharmacology, 10. SAGE Publications doi:10.1177/2045125320959560.

Cho, J., et al., 2018. Clozapine and all-cause mortality in treatment-resistant schizophrenia: a historical cohort study. Acta Psychiatrica Scandinavica, 139. Blackwell Publishing Ltd p. acps.12989. doi: doi:10.1111/acps.12989.

Chou, F.H.C., Tsai, K.Y., Chou, Y.M., 2013. The incidence and all-cause mortality of pneumonia in patients with schizophrenia: A nine-year follow-up study. J. Psychiatr. Res. 47 (4), 460–466. doi:10.1016/j.jpsychires.2012.12.007, Elsevier Ltd.

CRIS NLP Applications Library (2020) CRIS natural language processing, v1.1. Available at: https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/cris-natural-language-processing/.

Department of Health and Social Care, 2020. New UK-wide Methodology Agreed to Record COVID-19 Deaths - GOV.UK, at. Gov.Uk Available https://www.gov.uk/government/news/new-uk-wide-methodology-agreed-to-record-covid-19-deaths.

Govind, R., et al., 2020. Clozapine treatment and risk of COVID-19 infection: retrospective cohort study. Br. J. Psychiatry R. Coll. Psychiatr. 1–7. doi:10.1192/bjp.2020.151.

Haddad, P.M., 2013. Current use of second-generation antipsychotics may increase risk of pneumonia in people with schizophrenia. Evid. Based Ment. Health 16 (4), 109. doi:10.1136/eb-2013-101441.

Hayes, J.F., et al., 2017. Mortality gap for people with bipolar disorder and schizophrenia: UK-based cohort study 2000-2014. Br. J. Psychiatry R. Coll. Psychiatr. 175–181. doi:10.1192/bjp.bp.117.202606.

Hayes, R.D., et al., 2015. The effect of clozapine on premature mortality: an assessment of clinical monitoring and other potential confounders. Schizophr. Bull. 41 (3), 644–655. doi:10.1093/schbul/sbu120.

John, A., et al., 2018. Premature mortality among people with severe mental illness - New evidence from linked primary care data. Schizophr. Res. 199, 154–162. doi:10.1016/j.schres.2018.04.009, Elsevier B.V..

Kesserwani, J., et al., 2019. Risk of readmission in patients with schizophrenia and schizoaffective disorder newly prescribed clozapine. J. Psychopharmacol. 33 (4), 449–458. doi:10.1177/0269881118817387, SAGE Publications Ltd.

Kuo, C.-J., et al., 2013. Second-generation antipsychotic medications and risk of pneumonia in schizophrenia. Schizophr. Bull. 39 (3), 648–657. doi:10.1093/schbul/sbr202.

De Leon, J., Sanz, E.J., De las Cuevas, C., 2020. Data from the World Health Organization's pharmacovigilance database supports the prominent role of pneumonia in mortality associated with clozapine adverse drug reactions. Schizophr. Bull. 46 (1), 1–3. doi:10.1093/schbul/sbz093.

Liu, N.H., et al., 2017. Excess mortality in persons with severe mental disorders: a multilevel intervention framework and priorities for clinical practice, policy and research agendas. World Psychiatry 16 (1), 30–40. doi:10.1002/wps.20384, Blackwell Publishing Ltd.

Newcomer, J.W., 2005. Second-generation (atypical) antipsychotics and metabolic effects. In: CNS Drugs. Springer Nature, pp. 1–93. doi:10.2165/00023210-200519001-00001 19(Supplement 1).

Perera, G., et al., 2016. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data re-

source. BMJ Open 6 (3). doi:10.1136/BMJOPEN-2015-008721, BMJ Publishing Group.

Public Health England (2020) PHE reporting of COVID-19 deaths: technical summary, 12 August 2020. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/916035/RA_Technical_Summary_-_PHE_Data_Series_COVID_19_Deaths_20200812.pdf.

Seminog, O.O., Goldacre, M.J., 2013. 'Risk of pneumonia and pneumococcal disease in people with severe mental illness: English record linkage studies'. Thorax 68 (2), 171–176. doi:10.1136/thoraxjnl-2012-202480, BMJ Publishing Group.

Shen, T.C., et al., 2018. Risk of pleural empyema in patients with schizophrenia: A nationwide propensity-matched cohort study in Taiwan. BMJ Open 8 (7). doi:10.1136/bmjopen-2017-021187, BMJ Publishing Group.

Siskind, D., et al., 2016. Clozapine v. first- and second-generation antipsychotics in treatment-refractory schizophrenia: systematic review and meta-analysis. Br. J. Psychiatry 209 (5), 385–392. doi:10.1192/bjp.bp.115.177261, Royal College of Psychiatrists.

Siskind, D., et al., 2020. Consensus statement on the use of clozapine during the COVID-19 pandemic. J. Psychiat. Neurosci. JPN 45 (3), 2. doi:10.1503/jpn.200061, NLM (Medline).

Stewart, R., et al., 2009. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. BMC Psychiatry 9 (1), 51. doi:10.1186/1471-244X-9-51.

Stoecker, Z.R., et al., 2017. Clozapine usage increases the incidence of pneumonia compared with risperidone and the general population: A retrospective comparison of clozapine, risperidone, and the general population in a single hospital over 25 months. Int. Clin. Psychopharmacol. 32 (3), 155–160. doi:10.1097/YIC.0000000000000162, Lippincott Williams and Wilkins.

Tabary, M., et al., 2020. Pathologic features of COVID-19: A concise review. Pathology Research and Practice. Elsevier GmbH doi:10.1016/j.prp.2020.153097.

Vermeulen, J.M., et al., 2019. Clozapine and long-term mortality risk in patients with schizophrenia: a systematic review and meta-analysis of studies lasting 1.1-12.5 years. In: Schizophrenia Bulletin. Oxford University Press, pp. 315–329. doi:10.1093/schbul/sby052.

Vita, A., Barlati, S., 2021. The impact of the Covid-19 pandemic on patients with schizophrenia. European Neuropsychopharmacology. Eur Neuropsychopharmacol p. [Online ahead of print]. doi:10.1016/j.euroneuro.2021.08.003.

Wimberley, T., et al., 2017. Mortality and self-harm in association with clozapine in treatment-resistant schizophrenia. Am. J. Psychiatry 174 (10), 990–998. doi:10.1176/appi.ajp.2017.16091097, American Psychiatric Association.

World Health Organization, 1995. Physical status: the Use and Interpretation of anthropometry. Report of a WHO Expert Committee. World Health Organization technical report series, Switzerland doi:10.1002/(sici)1520-6300(1996)8:6⟨786::aid-ajhb11⟩3.0.co;2-i.

CHAPTER 7

7 DISCUSSION

7.1 SUMMARY OF KEY FINDINGS

In this thesis, four studies were performed in order to investigate clozapine health outcomes using electronic health records (EHR). Fully de-identified EHR data from South London and Maudsley NHS Foundation Trust (SLAM) was accessed via the Clinical Record Interactive Search (CRIS) system (Stewart *et al.*, 2009). All data used in this thesis was extracted from CRIS.

Chapter 3 demonstrated the challenges of working with EHR data for research. We performed a study to investigate variables that may be associated with clozapine-induced blood dyscrasia, and some potential predictors were even found. However, we also found that our cohort was not informative enough for this research question. EHR is a vast data resource that comes with almost no documentation. This was my first time working with EHR data. The knowledge and experience gained during this work played a crucial role in the study design of chapters 4, 5 and 6.

Chapter 4 is the first study (manuscript in preparation) to identify clozapine treatment start dates using clozapine blood monitoring data. Regular blood monitoring is a compulsory requirement for receiving clozapine treatment. Therefore, utilizing the date of blood tests as the basis to identify clozapine treatment dates has the potential to give us more reliable start and stop dates of the clozapine treatment periods. Using the identified start dates of

clozapine treatment, we investigated the change in risk of clozapine-induced blood dyscrasia over the clozapine treatment time. The findings of this chapter showed that there is a relatively high risk of blood dyscrasia is at the beginning of clozapine treatment, which is significantly reduced after 6 months of treatment. This contrast indicates the possibility of more than one biological mechanism being responsible for the blood dyscrasias occurring during clozapine treatment, but more work is required to confirm this theory. Secondly, this chapter also included findings on the risk of having a future blood dyscrasia event and shows that as clozapine treatment progresses, the risk-benefit ratio of monitoring changes significantly. This data has the potential to impact the future of clozapine monitoring towards reducing the monitoring even further as treatment progresses.

At the onset of the COVID-19 pandemic in the UK, I took the opportunity to investigate whether clozapine treatment was associated with an increased risk of COVID-19 infection (chapter 5). I followed up this study by investigating whether clozapine treatment was associated with an increased risk of adverse outcomes of COVID-19, namely COVID-related hospitalisation and intensive care treatment, and death (chapter 6).

In Chapter 5, I tested for associations between clozapine treatment and increased risk of COVID-19 in patients with schizophrenia-spectrum disorders. Of 6,309 patients in the cohort, 102 tested positive for COVID-19. Clozapine-treated patients showed an increased risk of COVID-19 compared with those who were on other antipsychotic medication (unadjusted HR=2.62 (95% CI 1.73 - 3.96), which was attenuated after adjusting for potential confounders, including clinical contact (adjusted hazard ratio HR=1.76, 95% CI 1.14 - 2.72). The findings provided support for the hypothesis that clozapine treatment is associated with an increased risk of COVID-19. Since the study was performed soon after the onset of the

COVID-19 pandemic in the UK, replicating the study in other samples and with current data is needed to confirm this association.

Chapter 6 was a follow-up study to investigate associations between clozapine treatment and increased risk of adverse outcomes of COVID-19. In our sample of patients with COVID-19 and schizophrenia-spectrum disorders, we found no evidence that clozapine treatment puts patients at increased risk of hospitalisation, intensive care treatment or death, compared to other antipsychotic treatments. The associations between clozapine and each outcome were tested using logistic regression models, adjusting for gender, age, ethnicity, neighbourhood deprivation, obesity, smoking status, diabetes, asthma, bronchitis and hypertension using propensity scores. In the unadjusted analysis, there was no significant association between clozapine and any of the outcomes and there remained no associations following adjusting for the confounding variables. Further research is needed in larger samples to confirm this.

7.2 STRENGTHS AND LIMITATIONS

A common strength of all studies reported in this thesis is that the studies were based on patients from SLAM. SLAM has a near-monopoly in providing all aspects of secondary mental health care to over 1.3 million people of four London boroughs (Lambeth, Southwark, Lewisham, and Croydon), thus creating an ideal research environment to confidently study the effects of antipsychotic drugs, such as clozapine, on patients in the UK.

Furthermore, all studies were based on data extracted from SLAM's EHR data. EHR data is comprised of valuable 'real-world' and 'real-time' information in large volumes. The statistical power and real-time nature of EHR help answer pressing research questions

quickly via retrospective studies. Working with EHR data eliminates the time-consuming step of patient recruitment needed for other types of studies such as an observational study (Perera *et al.*, 2009; Stewart, 2014). It is not feasible to perform an observational study on large cohorts, as close to real-time as possible and in rapid time. For example, within 5 months from the onset of the COVID-19 pandemic in the UK, we published a study that provided support for the hypothesis that clozapine treatment is associated with an increased risk of COVID-19. The study cohort included over 6,000 patients.

Moreover, we accessed SLAM's EHR data via Clinical Record Interactive Search (CRIS). The CRIS database not only contains SLAM's de-identified clinical records and makes them available for researchers, but also contains Natural Language Processing (NLP) data and linkage data.

The NLP data comes from custom-built algorithms that extract information from free-text clinical notes and store the results in a format that researchers can easily incorporate in their studies. A key strength of having the ready-made NLP results is it saves researchers probably hundreds of hours of work. Typically, EHR is made of structured and unstructured free-text data. The majority of the information resides in the free-text data but in order to use this information in research, one needs to either: (a) read the free-text clinical notes and extract the information manually, or (b) develop an NLP algorithm so that the data extraction can be performed computationally. Both these options are time-consuming and therefore it will not be feasible for a researcher to extract many variables this way. CRIS has a dedicated team that builds and maintains NLP algorithms. This gives researchers using CRIS a wide range of ready variables to design a study that has several covariates. Chapters

3, 5 and 6 included covariates that came from the results of NLP algorithms. These studies would have been less robust without the available NLP data.

The linkage data comes from CRIS securely linking with external clinical databases that can give researchers more information about SLAM's patients, information that could contribute to improving research studies. One example of such a database is the Zaponex Treatment Access System (ZTAS) database, which includes the clozapine blood monitoring data of clozapine patients treated at SLAM. The study in chapter 4 is based on the ZTAS data and it would not have been possible to do this study without the linkage to the ZTAS database.

The primary limitation of using EHR data for research is that it is a large volume of data that comes with limited documentation. There is a learning curve involved when beginning to design studies based on EHR data. To be able to successfully design a robust study, prior knowledge of how information is organised in the EHR and the best source for each type of information is needed. This prior knowledge generally comes after experience of working with the specific EHR data. For example, chapter 3 was my first time working with EHR data and this study encountered various limitations. However, the experience gained during chapter 3 played a crucial role in the rest of my thesis.

Another major limitation of EHR data is that its most useful and robust source of information is the free-text clinical notes. The free-text notes are the clinical narratives written by physicians, nurses and other healthcare providers. They hold detailed records of almost every interaction the hospital has had with each patient, such as via face-to-face consultations, telephone calls, and email correspondence. The only ways to extract information for the clinical narratives are via reading manually or via complex

computational solutions, for example, NLP. Fortunately, CRIS contains ready-made results of several NLP algorithms, some of which were even used in the thesis. Unfortunately, not all variables that a researcher needs for their study are catered for by the existing NLP algorithms. For example, in chapter 6, I had to manually read all relevant clinical notes to extract information on all the outcomes and some covariates needed for the study.

A third major limitation of EHR data is the problem of missing data. In this thesis, missing data was a problem in all the research chapters and was handled differently in each chapter. In chapter 3, there were 9 variables that contained more than 25% missing data. Since logistic regression was used in the statistical analysis, a conservative threshold of maximum 25% missing data was set for including variables in the statistical analysis. Performing logistic regression on a dataset that includes missing values, results in a reduction in sample size and power. In chapter 4, missing data had a significant impact on the results. The algorithm we developed for identifying clozapine treatment start dates was not able to identify the start dates of 40% of the treatment periods. This was due to the unavailability of the relevant data from that time. We performed the statistical analysis on the start dates that the algorithm was able to identify. In chapter 5, there were 3 potential confounders containing missing data: ethnicity, smoking status, and Body Mass Index. Here, we first analysed the entire cohort using only variables with no missing data. Then, in a complete case analysis, we re-ran the same analyses on individuals who did not have any missing data. Since the results were very similar, we present results from the complete case analysis and included the results of the whole-cohort analysis in the supplementary material. In chapter 6, there were 7 potential confounders containing missing data: ethnicity, smoking status, obesity status, diabetes, asthma, bronchitis, and hypertension. In this chapter, all

missing values were replaced by values that were manually extracted by going through the free-text clinical notes of the EHR. This was the most time-consuming part of this analysis, but it was required because logistic regression was used in the statistical analysis.

Surveillance bias is potentially another limitation seen in this thesis. Surveillance bias refers to the non-random type of information bias summarised by this phrase: “the more you look, the more you find.” (Pierce et al., 2008). Surveillance bias can occur when certain patients are followed up more closely and therefore have more diagnostics tests performed on them compared to others. Therefore, one group can appear to have a higher proportion of a diagnosis but that is a data artifact caused by surveillance bias. In chapter 5, we compared clozapine-treated patients with patients on any non-clozapine antipsychotic(s). The outcome measure was COVID-19 infection. It cannot be ruled out that since clozapine patients are more closely monitored, this group might have more COVID-19 tests performed. One way of checking for potential surveillance bias would have been to perform a parallel analysis with the outcome measure as getting a COVID-19 test (irrespective of the positive or negative results) to check if there was an association.

7.3 Implications and Future Work

To our knowledge, chapter 5 incorporates the first study to suggest that patients on clozapine treatment are at higher risk of infection by COVID-19. This is consistent with previous research demonstrating that patients treated with clozapine have higher rates of infection and pneumonia than patients on other antipsychotics (Walker *et al.*, 1997; Copeland *et al.*, 2007; Kuo *et al.*, 2013; Hung *et al.*, 2016; Leung *et al.*, 2017; Stoecker *et al.*, 2017). However, this study was based on the data from March 2020 to May 2020, which

was only the first 3 months of the COVID-19 pandemic in the UK. Replicating this study in other samples and with current data is needed to firmly establish this confirmation. Chapter 6, the follow-up study, found no associations between clozapine treatment and an increased risk of adverse outcomes of COVID-19, but was underpowered. Therefore, the next step is to replicate both studies in other settings to confirm the findings.

A study by Anna Ohlis and colleagues investigated the associations between clozapine treatment and an increased risk of adverse outcomes of COVID-19 in a large Swedish population study (*N* = 8,233) (Ohlis *et al.*, 2022). Their cohort comprised all adult residents in the Stockholm Region who had a psychotic disorder diagnosis and were receiving antipsychotic treatment. The exposed group were patients on clozapine treatment, and they were compared with a group of patients who were on non-clozapine antipsychotics. There were 3 outcome measures: inpatient care, intensive care treatment or death due to COVID-19 infection. Their results showed no statistically significant differences between the two groups of patients for any of the outcome measures. These finding are consistent with the findings in study in chapter 6. Together, the results add support to the current clozapine treatment guidelines which is to continue clozapine treatment during the current COVID-19 pandemic with careful monitoring (Siskind, Honer, *et al.*, 2020).

Another study by Katlyn Nemani and colleagues investigated the association between a diagnosis of schizophrenia spectrum, mood, or anxiety disorders and an increased risk of mortality in patients with COVID-19 (Nemani *et al.*, 2021). Their cohort comprised 7,348 adults from a New York health system who had laboratory-confirmed COVID-19. The exposed group were patients with schizophrenia spectrum, mood, or anxiety disorders diagnosis and they were compared with a reference group of patients who had no

psychiatric disorders. The outcome measure was mortality which was defined as death within 45 days following a positive COVID-19 test result. Their results showed that mood and anxiety disorders were not associated with an increased risk of mortality. However, diagnosis of a schizophrenia spectrum disorder was significantly associated with mortality (OR 2.67; 95% CI, 1.48-4.80). Unlike the study in chapter 6, which had schizophrenia spectrum disorder as an inclusion criterion, this study had schizophrenia spectrum disorder as the exposure of interest. Together, the studies help us conclude that compared to individuals with no psychiatric disorders, patients with schizophrenia spectrum disorder are at an increased risk of COVID-19 related mortality. However, amongst the patients with schizophrenia spectrum disorder, being on clozapine treatment does not increase this risk. This also adds support to the current clozapine treatment guidelines which is to continue clozapine treatment during the current COVID-19 pandemic with careful monitoring (Siskind, Honer, *et al.*, 2020).

A recent study found that clozapine-treated patients display clinical patterns that resemble primary immunodeficiency common variable immunodeficiency (CVID), which reverses with clozapine discontinuation (Ponsford *et al.*, 2020). CVID is an immune deficiency disease characterized by low levels of protective antibodies and an increased risk of infections. Another recent study that involved antibody profiling found reduced antibody counts in clozapine treated patients (Jernbom Falk *et al.*, 2021). This poses an important clinical question of whether clozapine treatment affects the efficacy of the COVID-19 vaccines. To test if clozapine treatment reduces COVID-19 vaccine efficacy, a study can be designed based on the recent Omicron wave in the UK when the majority of the population were vaccinated but there were still some unvaccinated people. The analysis will involve

calculating hazard ratios for COVID-19-positive status in clozapine-treated patients compared with those treated with other antipsychotics. Separate analysis can be performed for the vaccinated and the unvaccinated populations and results compared. If the findings suggest that clozapine treatment is associated with reduced efficacy of COVID-19 vaccine, and this association is replicated and becomes firmly established, then clinicians and patients will need to weigh up the risk of reduced immunity against COVID-19 infection against the risk of psychotic relapse if clozapine is discontinued.

Chapter 4 was a study investigating the frequency of neutropenia using the clozapine blood test monitoring data. The next step for this analysis is to distinguish between the clozapine-induced neutropenia and those that were caused by reasons unrelated to the clozapine treatment. This can be done by thoroughly reading EHR free-text fields of the 75 patients who had neutropenia during their clozapine treatment. The results of this chapter can also be used to improve patient care via Population Health Management.

7.4 Population Health Management: the future of patient care

Population Health Management (PHM) is an innovative data-driven approach to improving patient care by facilitating the identification of the specific populations who need to be prioritised for a specific health or care need (Embuldeniya *et al.*, 2021; NHS England, 2021). It is based on using existing clinical data (historic and present-day) to predict current and future trends in the health and care needs of sub-populations with the aim to provide tailored support for individuals. Digital PHM platforms have the potential to help the frontline teams quickly identify factors that are causing poor outcomes in specific

populations. One such digital tool is VIEWER (Visualisation & Interaction With Electronic Records) (Codling *et al.*, 2021).

VIEWER is a newly developed PHM dashboarding tool that is created to help the frontline teams identify specific groups of people with psychosis who are at risk of poor outcomes. The motivation to develop VIEWER was driven by the aim to help identify factors that cause poor health outcomes in patients with psychosis. Patients with psychosis have, on average, poorer physical health than the rest of the population (Kim *et al.*, 2019). Moreover, patients with psychosis who have treatment-resistant schizophrenia (TRS) and should be offered clozapine are not often identified in the early stages of psychosis. One of the aims of VIEWER is to help in the early recognition of TRS.

Figure 7.a and Figure 7.b shows screenshots from VIEWER presenting plots of bed usage, antipsychotic prescribing and a heatmap of patients with psychosis. All visualisations (graphs, heatmaps, charts) are dynamically generated in real-time based on the current data in the EHR.

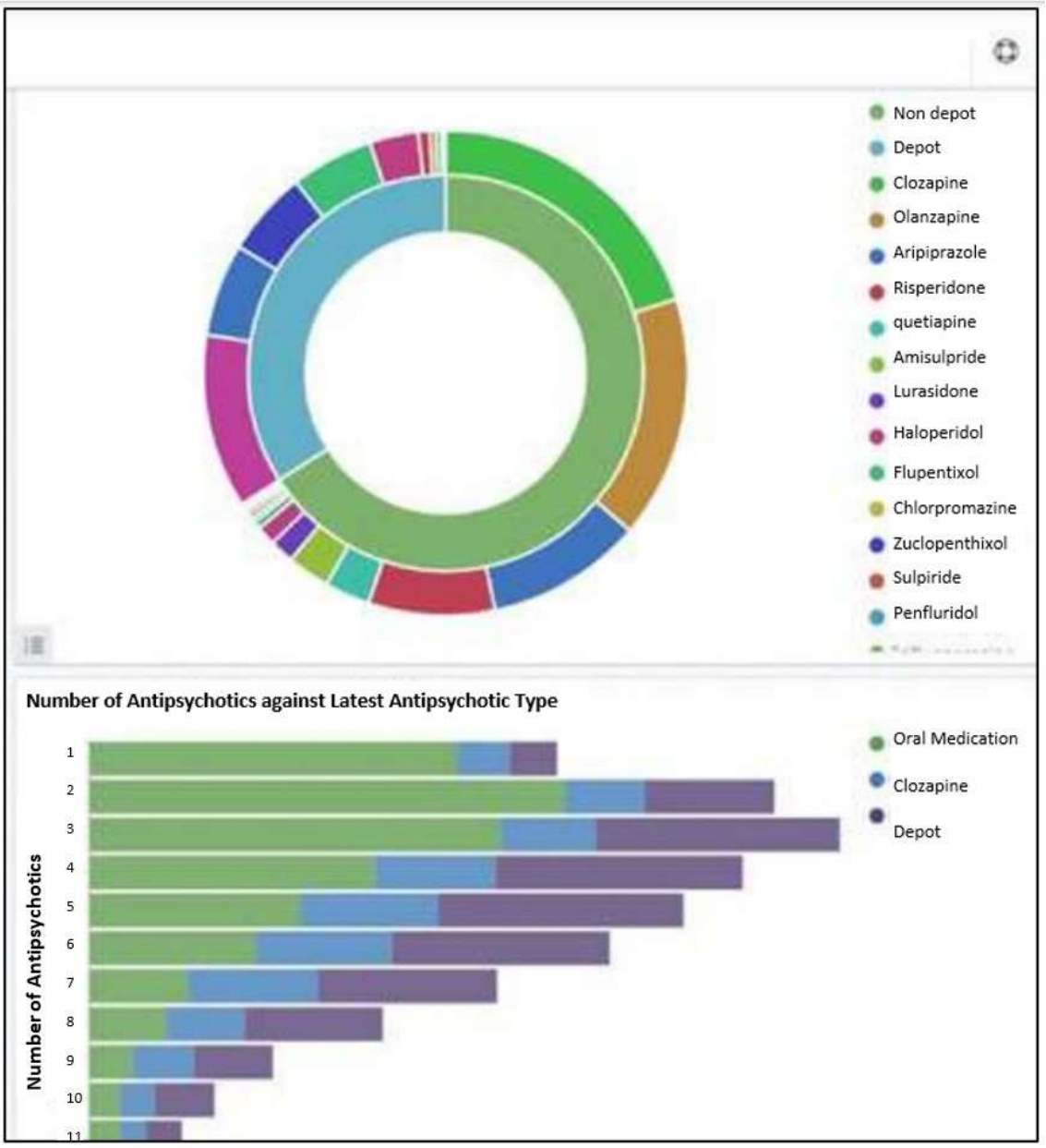
VIEWER was developed as a collaborative project between a team of clinicians, service users (patients), carers, clinical researchers, clinical informaticians and computer scientists at SLAM. It utilizes artificial intelligence-based NLP technologies to decipher population-level trends hidden within the EHR data and presents them to the frontline team as interactive visualisations (graphs, heatmaps, charts). All visualisations are dynamically generated and are interactive, thus allowing the user to manipulate the filters based on their own population of interest and define their own “at risk” group. The sub-categories of all visualisations are clickable and zoom in to a new page with visualisation on the newly

selected sub-population. The users can click through all the way down to a specific individual and then click through to read their EHR data.

Figure 7.a: Screenshot of the interactive VIEWER dashboard

The above screenshot shows two dynamically generated plots of: (i) Number of psychosis patients vs bed days over last 6 months (ii) heatmaps of neighbourhood of psychosis patients (Codling et al., 2021)

Figure 7.b: Screenshot of the interactive VIEWER dashboard



The above screenshots shows two dynamically generated plots of: (i) proportion of patients on each antipsychotic medication (ii) Polypharmacy information, showing the number of antipsychotics against the type of the latest antipsychotic the patient is taking (Codling et al., 2021)

Figure 7.a and Figure 7.b are a is a small subset of different types of information available on VIEWER; it contains multi-dimensional information about patients, such as their diagnoses, medications, physical health (e.g., BMI), and service usage information. Almost all data presented in VIEWER was originally embedded within the free-text clinical notes and it was

through data-driven algorithms that these data were extracted and analysed to reveal health and care trends that were difficult to see before. Visualisations of these results are now made available to clinicians via VIEWER to improve patient care.

In Chapter 4, I studied the patterns in the clozapine blood test monitoring data and identified the dates of when patients started and stopped their clozapine treatments. Regular blood monitoring is a compulsory requirement for receiving clozapine treatment, therefore treatment dates identified from this data has the potential to be more reliable than treatment dates extracted from free-text fields using NLP, which is the current source of clozapine treatment dates data in CRIS.

In addition, the data from Chapter 4 also includes information of the time from the start of a clozapine treatment to the first red flag in the blood monitoring results. The red flags indicate the risk of blood dyscrasia, the side effect of clozapine that warrants the clozapine blood monitoring. The number of days from a clozapine treatment to the first red flag data can be presented as a histogram similar to Bed Days in Figure 7.a. Currently, there are no clinical ways to predict who is at risk for clozapine-induced blood dyscrasia. This data will help clinicians visualise plots on population levels as well as the sub-population levels, based on the demographics of their patients. The plots can also be used to evaluate the likelihood of a future red result based on how far into the treatment a patient is. Thus, this data has the potential to improve patient care by helping clinicians make better-informed decisions about whether to continue clozapine, whether to restart it if it had to be stopped.

PHM tools like VIEWER aim to use various data-driven approaches to enhance a clinician's ability to provide optimal care to each patient, and thereby improving the clinical outcomes of the population in general.

7.5 CONCLUSION

In conclusion, EHR data is a valuable resource for performing clinical research, including investigating clozapine health outcomes. Using EHR data, I was able to identify clozapine treatment start and stop dates. This has the potential to be integrated into VIEWER and help improve patient care. Also, using EHR data, I performed a study that provided support for the hypothesis that clozapine treatment is associated with an increased risk of COVID-19. This study was published within 5 months of the onset of the COVID-19 pandemic. EHR data provides the platform to perform studies on relatively large cohorts in a shorter span of time, compared to conventional research studies which require patient recruitment. Lastly, using EHR data, I performed a follow-up study that showed that even though the clozapine treatment may be associated with an increased risk of COVID-19, we found no evidence that clozapine treatment puts patients at increased risk of adverse outcomes of COVID-19, namely hospitalisation, intensive care treatment or death.

CHAPTER 8

8 REFERENCES

Ambinder, E. P. (2005) 'Electronic Health Records', *Journal of Oncology Practice*. American Society of Clinical Oncology, 1(2), p. 57. doi: 10.1200/JOP.2005.1.2.57.

Amsler, H. A. *et al.* (1977) 'Agranulocytosis in patients treated with clozapine: A STUDY OF THE FINNISH EPIDEMIC', *Acta Psychiatrica Scandinavica*. John Wiley & Sons, Ltd, 56(4), pp. 241–248. doi: 10.1111/j.1600-0447.1977.tb00224.x.

Andreasen, N. C. (1984) 'Scale for the Assessment of Positive Symptoms', *Medicine*, 17 (2), pp. 173–180.

Andreasen, N. C. (1989) 'The Scale for the Assessment of Negative Symptoms (SANS): Conceptual and theoretical foundations', *British Journal of Psychiatry*. Cambridge University Press, 155(NOV. SUPPL. 7), pp. 49–52. doi: 10.1192/s0007125000291496.

Bagattini, F. *et al.* (2019) 'A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records', *BMC Medical Informatics and Decision Making*. BMC Med Inform Decis Mak, 19(1). doi: 10.1186/s12911-018-0717-4.

Barnes, T. R. E. and Paton, C. (2011) 'Antipsychotic polypharmacy in Schizophrenia: Benefits and risks', *CNS Drugs*. CNS Drugs, 25(5), pp. 383–399. doi: 10.2165/11587810-000000000-00000.

Bloor, R. N. (1995) 'Setting up a psychiatric case register', *Advances in Psychiatric Treatment*, 1(3), pp. 86–91. doi: 10.1192/apt.1.3.86.

BNF (2020) *BNF: British National Formulary - NICE*. BMJ Group and Pharmaceutical Press.

Boyd A, Thomas R, M. J. (2018) 'NHS Numbers and their management systems.'

Caldicott, F. (2013) 'Information: To share or not to share', *The Information Governance Review Department of Health, Crown copyright*.

Capurro, D. *et al.* (2014) 'Availability of Structured and Unstructured Clinical Data for Comparative Effectiveness Research and Quality Improvement: A Multi-Site Assessment', *eGEMs (Generating Evidence & Methods to improve patient outcomes)*. Ubiquity Press, 2(1), p. 11. doi: 10.13063/2327-9214.1079.

Chakos, M. *et al.* (2001) 'Effectiveness of Second-Generation Antipsychotics in Patients With Treatment-Resistant Schizophrenia: A Review and Meta-Analysis of Randomized Trials', *American Journal of Psychiatry*, 158(4), pp. 518–526. doi: 10.1176/appi.ajp.158.4.518.

Chang, C. K. *et al.* (2011) 'Life expectancy at birth for people with serious mental illness and other major disorders from a secondary mental health care case register in London', *PLoS ONE*. PLoS One, 6(5). doi: 10.1371/journal.pone.0019590.

Che, Z. *et al.* (2018) 'Recurrent Neural Networks for Multivariate Time Series with Missing Values', *Scientific Reports*. Nature Publishing Group, 8(1). doi: 10.1038/s41598-018-24271-9.

Chen, D. *et al.* (2021) 'Early Detection of Post-Surgical Complications using Time-series Electronic Health Records.', *AMIA ... Annual Symposium proceedings. AMIA Symposium*. AMIA Annu Symp Proc, 2021, pp. 152–160.

Chen, S. Y. *et al.* (2019) 'Treatment Strategies for Clozapine-Induced Sialorrhea: A Systematic Review and Meta-analysis', *CNS Drugs*. CNS Drugs, 33(3), pp. 225–238. doi: 10.1007/s40263-019-00612-8.

Cheng, Y. *et al.* (2016) 'Risk prediction with electronic health records: A deep learning approach', *16th SIAM International Conference on Data Mining 2016, SDM 2016*. Society for Industrial and Applied Mathematics Publications, pp. 432–440. doi: 10.1137/1.9781611974348.49.

Codling, D. *et al.* (2021) 'VEIWER: a Digital tool for visualising data in mental health records VIEWER: a Digital tool for visualising data in mental health records'. doi: 10.21203/RS.3.RS-955124/V1.

Copeland, L. A. *et al.* (2007) 'Pulmonary disease among inpatient decedents: Impact of schizophrenia', *Progress in neuro-psychopharmacology & biological psychiatry*. Prog Neuropsychopharmacol Biol Psychiatry, 31(3), pp. 720–726. doi: 10.1016/J.PNPBP.2007.01.008.

Correll, C. U. *et al.* (2007) 'Does antipsychotic polypharmacy increase the risk for metabolic syndrome?', *Schizophrenia Research*. Schizophr Res, 89(1–3), pp. 91–100. doi: 10.1016/j.schres.2006.08.017.

Crane, J. and Bunn, S. (2016) 'Government Proposals for Health Records'.

Crilly, J. (2007) 'The history of clozapine and its emergence in the US market: A review and analysis', *History of Psychiatry*. Hist Psychiatry, 18(1), pp. 39–60. doi: 10.1177/0957154X07070335.

CRIS NLP Applications Library (2020) *CRIS Natural Language Processing*, *v1.1*. Available at: https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/cris-natural-language-processing/.

Dollfus, S. and Lyne, J. (2017) 'Negative symptoms: History of the concept and their position in diagnosis of schizophrenia', *Schizophrenia Research*. Schizophr Res, 186, pp. 3–7. doi: 10.1016/j.schres.2016.06.024.

Downs, J. M. *et al.* (2019) 'An approach to linking education, social care and electronic health records for children and young people in South London: A linkage study of child and adolescent

mental health service data', *BMJ Open*. BMJ Publishing Group, 9(1). doi: 10.1136/bmjopen-2018-024355.

Ekblom, B. and Haggstrom, J. E. (1974) *Clozapine (Leponex) compared with chlorpromazine: a double blind evaluation of pharmacological and clinical properties*, *Current Therapeutic Research - Clinical and Experimental*. Available at: https://psycnet.apa.org/record/1975-05854-001 (Accessed: 10 September 2021).

Embuldeniya, G. *et al.* (2021) 'The beginnings of health system transformation: How Ontario Health Teams are implementing change in the context of uncertainty', *Health Policy*. Elsevier, 125(12), pp. 1543–1549. doi: 10.1016/j.healthpol.2021.10.005.

Every-Palmer, S. *et al.* (2016) 'Clozapine-treated Patients Have Marked Gastrointestinal Hypomotility, the Probable Basis of Life-threatening Gastrointestinal Complications: A Cross Sectional Study', *EBioMedicine*. Elsevier B.V., 5, pp. 125–134. doi: 10.1016/j.ebiom.2016.02.020.

Fischer Cornelssen, K. A. and Ferner, U. J. (1976) 'An example of European multicenter trials: multispectral analysis of clozapine', *Psychopharmacology Bulletin*, 12(2), pp. 34–39.

Ford, E. *et al.* (2013) 'Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: What information is hidden in free text?', *BMC Medical Research Methodology*. BioMed Central, 13(1), pp. 1–12. doi: 10.1186/1471-2288-13-105.

GBD 2017 Disease and Injury Incidence and Prevalence Collaborators (2018) 'Global, regional, and national incidence, prevalence, and years lived with disability for 354 Diseases and Injuries for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017', *The Lancet*. Lancet Publishing Group, 392(10159), pp. 1789–1858. doi: 10.1016/S0140-6736(18)32279-7.

Gerlach, J. *et al.* (1974) 'Clozapine and Haloperidol in a Single-Blind Cross-Over Trial: Therapeutic and Biochemical Aspects in the Treatment of Schizophrenia', *Acta Psychiatrica Scandinavica*. Acta Psychiatr Scand, 50(4), pp. 410–424. doi: 10.1111/j.1600-0447.1974.tb09706.x.

Haddy, T. B., Rana, S. R. and Castro, O. (1999) 'Benign ethnic neutropenia: What is a normal absolute neutrophil count?', *Journal of Laboratory and Clinical Medicine*. J Lab Clin Med, 133(1), pp. 15–22. doi: 10.1053/lc.1999.v133.a94931.

Hägg, S., Spigset, O. and Söderström, T. G. (2000) 'Association of venous thromboembolism and clozapine', *Lancet*. Lancet, 355(9210), pp. 1155–1156. doi: 10.1016/S0140-6736(00)02066-3.

Hamann, J. *et al.* (2005) 'How do psychiatrists choose among different antipsychotics?', *European Journal of Clinical Pharmacology*. Eur J Clin Pharmacol, 61(11), pp. 851–854. doi: 10.1007/s00228-005-0041-7.

Harel, O. *et al.* (2018) 'Multiple Imputation for Incomplete Data in Epidemiologic Studies', *American Journal of Epidemiology*. Am J Epidemiol, 187(3), pp. 576–584. doi: 10.1093/aje/kwx349.

Hippius, H. (1989) 'The history of clozapine', *Psychopharmacology*. Springer-Verlag, 99(1 Supplement), pp. 3–5. doi: 10.1007/BF00442551.

Hippius, H. (1999) 'A historical perspective of clozapine', *Journal of Clinical Psychiatry*. Physicians Postgraduate Press, Inc., 60(SUPPL. 12), pp. 22–23.

Howes, O. D. *et al.* (2017) 'Treatment-ResistantSchizophrenia: TreatmentResponse and Resistance in Psychosis (TRRIP) Working Group Consensus Guidelines on Diagnosis and Terminology', *American Journal of Psychiatry*. Am J Psychiatry, 174(3), pp. 216–229. doi: 10.1176/appi.ajp.2016.16050503.

Howes, O. D. and Murray, R. M. (2014) 'Schizophrenia: An integrated sociodevelopmental-cognitive model', *The Lancet*. Elsevier, 383(9929), pp. 1677–1687. doi: 10.1016/S0140-6736(13)62036-X.

Hung, G. C. L. *et al.* (2016) 'Antipsychotic reexposure and recurrent pneumonia in schizophrenia: a nested case-control study', *The Journal of clinical psychiatry*. J Clin Psychiatry, 77(1), pp. 60–66. doi: 10.4088/JCP.14M09301.

Idänpään-Heikkilä, J. *et al.* (1975) 'Clozapine and agranulocytosis', *The Lancet*. Lancet, 306(7935), p. 611. doi: 10.1016/S0140-6736(75)90206-8.

Iqbal, E. *et al.* (2017) 'ADEPt, a semantically-enriched pipeline for extracting adverse drug events from free-text electronic health records', *PLoS ONE*. Edited by T. Groza. Public Library of Science, 12(11), p. e0187121. doi: 10.1371/journal.pone.0187121.

Iqbal, E. *et al.* (2020) 'The side effect profile of Clozapine in real world data of three large mental health hospitals', *PLoS ONE*. Edited by V. De Luca. Public Library of Science, 15(12 December), p. e0243437. doi: 10.1371/journal.pone.0243437.

Irving, J. *et al.* (2021) 'Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk', *Schizophrenia Bulletin*. Schizophr Bull, 47(2), pp. 405–414. doi: 10.1093/schbul/sbaa126.

Jackson, R. G. *et al.* (2017) 'Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project.', *BMJ open*. British Medical Journal Publishing Group, 7(1), p. e012012. doi: 10.1136/bmjopen-2016-012012.

Jayasinghe, L. *et al.* (2020) 'Clinician-recalled quoted speech in electronic health records and risk of suicide attempt: A case-crossover study', *BMJ Open*. British Medical Journal Publishing Group, 10(4), p. e036186. doi: 10.1136/bmjopen-2019-036186.

Jernbom Falk, A. *et al.* (2021) 'Autoantibody profiles associated with clinical features in psychotic

disorders', *Translational Psychiatry*. Nature Publishing Group, 11(1). doi: 10.1038/s41398-021-01596-0.

Joling, K. J. *et al.* (2011) 'Do GPs' medical records demonstrate a good recognition of depression? A new perspective on case extraction', *Journal of Affective Disorders*. Elsevier, 133(3), pp. 522–527. doi: 10.1016/j.jad.2011.05.001.

Kadra, G. *et al.* (2018) 'Antipsychotic polypharmacy prescribing and risk of hospital readmission', *Psychopharmacology*. Psychopharmacology (Berl), 235(1), pp. 281–289. doi: 10.1007/s00213-017-4767-6.

Kahn, R. S. *et al.* (2008) 'Effectiveness of antipsychotic drugs in first-episode schizophrenia and schizophreniform disorder: an open randomised clinical trial', *The Lancet*. Lancet, 371(9618), pp. 1085–1097. doi: 10.1016/S0140-6736(08)60486-9.

Kane, J. *et al.* (1988) 'Clozapine for the Treatment-Resistant Schizophrenic: A Double-blind Comparison With Chlorpromazine', *Archives of General Psychiatry*. Arch Gen Psychiatry, 45(9), pp. 789–796. doi: 10.1001/archpsyc.1988.01800330013001.

Kay, S. R., Fiszbein, A. and Opler, L. A. (1987) 'The positive and negative syndrome scale (PANSS) for schizophrenia', *Schizophrenia Bulletin*. Schizophr Bull, 13(2), pp. 261–276. doi: 10.1093/schbul/13.2.261.

Kennedy, J. L. *et al.* (2014) 'The social and economic burden of treatment-resistant schizophrenia: A systematic literature review', *International Clinical Psychopharmacology*. Int Clin Psychopharmacol, 29(2), pp. 63–76. doi: 10.1097/YIC.0b013e32836508e6.

Khokhar, J. Y. *et al.* (2018) 'Unique Effects of Clozapine: A Pharmacological Perspective', *Advances in Pharmacology*. NIH Public Access, 82, pp. 137–162. doi: 10.1016/bs.apha.2017.09.009.

Kim, S. W. *et al.* (2019) 'Physical health literacy and health-related behaviors in patients with psychosis', *Clinical Psychopharmacology and Neuroscience*. Korean College of Neuropsychopharmacology, 17(2), pp. 279–287. doi: 10.9758/cpn.2019.17.2.279.

Kuo, C.-J. *et al.* (2013) 'Second-generation antipsychotic medications and risk of pneumonia in schizophrenia.', *Schizophrenia bulletin*, 39(3), pp. 648–57. doi: 10.1093/schbul/sbr202.

Lehman, A. F. *et al.* (2004) 'Practice Guideline for the Treatment of Partients with Schizophrenia, Second Edition', *American Journal of Psychiatry*, 161(2 SUPPL.), pp. 1–56. doi: 10.1176/appi.books.9780890423363.45859.

Leucht, S. *et al.* (2012) 'Antipsychotic drugs versus placebo for relapse prevention in schizophrenia: A systematic review and meta-analysis', *The Lancet*. Lancet, 379(9831), pp. 2063–2071. doi: 10.1016/S0140-6736(12)60239-6.

Leung, J. G. *et al.* (2017) 'Characterization of Admission Types in Medically Hospitalized Patients Prescribed Clozapine', *Psychosomatics*. Psychosomatics, 58(2), pp. 164–172. doi: 10.1016/J.PSYM.2016.11.013.

Li, X. H. *et al.* (2020) 'The prevalence of agranulocytosis and related death in clozapine-treated patients: A comprehensive meta-analysis of observational studies', *Psychological Medicine*. Cambridge University Press, 50(4), pp. 583–594. doi: 10.1017/S0033291719000369.

Liddle, P. F. (1987) 'The Symptoms of Chronic Schizophrenia: A Re-examination of the Positive-Negative Dichotomy', *The British Journal of Psychiatry*. Cambridge University Press, 151(2), pp. 145–151. doi: 10.1192/BJP.151.2.145.

Lindenmayer, J. P. (2000) 'Treatment refractory schizophrenia', *Psychiatric Quarterly*. Psychiatr Q, 71(4), pp. 373–384. doi: 10.1023/A:1004640408501.

López-Muñoz, F. *et al.* (2005) 'History of the discovery and clinical introduction of chlorpromazine', *Annals of Clinical Psychiatry*. Ann Clin Psychiatry, 17(3), pp. 113–135. doi: 10.1080/10401230591002002.

Mansournia, M. A. and Altman, D. G. (2016) 'Inverse probability weighting', *BMJ (Online)*. BMJ, 352. doi: 10.1136/bmj.i189.

Marteene, W. *et al.* (2019) 'Strategies to counter antipsychotic-associated weight gain in patients with schizophrenia', *Expert Opinion on Drug Safety*. Expert Opin Drug Saf, 18(12), pp. 1149–1160. doi: 10.1080/14740338.2019.1674809.

McEvoy, J. P. *et al.* (2006) 'Effectiveness of clozapine versus olanzapine, quetiapine, and risperidone in patients with chronic schizophrenia who did not respond to prior atypical antipsychotic treatment', *American Journal of Psychiatry*. American Psychiatric Association, 163(4), pp. 600–610. doi: 10.1176/ajp.2006.163.4.600.

Meltzer, H. Y. (1992) 'Treatment of the neuroleptic-nonresponsive schizophrenic patient.', *Schizophrenia bulletin*, 18(3), pp. 515–42.

Meltzer, H. Y. (2012) 'Clozapine: Balancing safety with superior antipsychotic efficacy', *Clinical Schizophrenia and Related Psychoses*, 6(3), pp. 134–144. doi: 10.3371/CSRP.6.3.5.

Meyer, N. *et al.* (2015) 'Optimizing outcomes in clozapine rechallenge following neutropenia: A cohort analysis', *Journal of Clinical Psychiatry*. J Clin Psychiatry, 76(11), pp. e1410–e1416. doi: 10.4088/JCP.14m09326.

MHRA (2005) *Antipsychotic medicines - Medicines and Healthcare Products Regulatory Agency - UK Government*.

MHRA (2020) *Yellow Card Scheme - Medicines and Healthcare products Regulatory Agency*. Available

at: https://yellowcard.mhra.gov.uk/the-yellow-card-scheme/ (Accessed: 29 June 2021).

Morrison, M. (2020) 'Research using free text data in medical records could benefit from dynamic consent and other tools for responsible governance', *Journal of Medical Ethics*. Institute of Medical Ethics, 46(6), pp. 380–381. doi: 10.1136/medethics-2020-106189.

Mortimer, A. M. *et al.* (2010) 'Clozapine for treatment-resistant schizophrenia: National Institute of Clinical Excellence (NICE) guidance in the real world', *Clinical Schizophrenia and Related Psychoses*. Clin Schizophr Relat Psychoses, 4(1), pp. 49–55. doi: 10.3371/CSRP.4.1.4.

Munro, J. *et al.* (1999) 'Active monitoring of 12760 clozapine recipients in the UK and Ireland: Beyond pharmacovigilance', *British Journal of Psychiatry*. Br J Psychiatry, 175(DEC.), pp. 576–580. doi: 10.1192/bjp.175.6.576.

Myles, N. *et al.* (2018) 'Meta-analysis examining the epidemiology of clozapine-associated neutropenia', *Acta Psychiatrica Scandinavica*. Blackwell Publishing Ltd, 138(2), pp. 101–109. doi: 10.1111/acps.12898.

National Institute For Clinical Excellence (2014) *Schizophrenia: The NICE guideline on core interventions in the treatment and management of schizophrenia in primary and secondary care; National Clinical Practice Guidelines Number CG82*. London.

Nemani, K. *et al.* (2021) 'Association of Psychiatric Disorders with Mortality among Patients with COVID-19', *JAMA Psychiatry*. American Medical Association, 78(4), pp. 380–386. doi: 10.1001/jamapsychiatry.2020.4442.

NHS England (2021) *NHS England » Population Health and the Population Health Management Programme*. Available at: https://www.england.nhs.uk/integratedcare/what-is-integrated-care/phm/ (Accessed: 21 December 2021).

NICE (2014) *Psychosis and schizophrenia in adults: prevention and management | Guidance and guidelines | NICE*. NICE.

NICE (2021) *the national institute for health and care excellence - NICE*. Available at: https://www.nice.org.uk/ (Accessed: 8 November 2021).

Nielsen, J. *et al.* (2011) 'Optimizing clozapine treatment', *Acta Psychiatrica Scandinavica*. Acta Psychiatr Scand, 123(6), pp. 411–422. doi: 10.1111/j.1600-0447.2011.01710.x.

Nielsen, J. *et al.* (2016) 'Worldwide differences in regulations of clozapine use', *CNS Drugs*. CNS Drugs, 30(2), pp. 149–161. doi: 10.1007/s40263-016-0311-1.

Ohlis, A. *et al.* (2022) 'Clozapine treatment and risk of severe COVID-19 infection', *Acta Psychiatrica Scandinavica*. Wiley-Blackwell, 145(1), pp. 79–85. doi: 10.1111/ACPS.13379.

Olfson, M. *et al.* (2016) 'Clozapine for schizophrenia: State variation in evidence-based practice', *Psychiatric Services*. Psychiatr Serv, 67(2), p. 152. doi: 10.1176/appi.ps.201500324.

Oloyede, E. *et al.* (2021) 'There Is Life After the UK Clozapine Central Non-Rechallenge Database', *Schizophrenia Bulletin*. Oxford Academic, 47(4), pp. 1088–1098. doi: 10.1093/schbul/sbab006.

Overall, J. E. and Gorham, D. R. (1962) 'The Brief Psychiatric Rating Scale', *Psychological Reports*. SAGE PublicationsSage CA: Los Angeles, CA, 10(3), pp. 799–812. doi: 10.2466/pr0.1962.10.3.799.

Pedersen, J. S. *et al.* (2021) 'Deep learning detects and visualizes bleeding events in electronic health records', *Research and Practice in Thrombosis and Haemostasis*. Blackwell Publishing Ltd, 5(4). doi: 10.1002/rth2.12505.

Perera, G. *et al.* (2009) 'The psychiatric case register: Noble past, challenging present, but exciting future', *British Journal of Psychiatry*. Cambridge University Press, 195(3), pp. 191–193. doi:

10.1192/bjp.bp.109.068452.

Perera, G. *et al.* (2016) 'Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource', *BMJ Open*. BMJ Publishing Group, 6(3). doi: 10.1136/BMJOPEN-2015-008721.

Perkins, N. J. *et al.* (2018) 'Principled Approaches to Missing Data in Epidemiologic Studies', *American Journal of Epidemiology*. Am J Epidemiol, 187(3), pp. 568–575. doi: 10.1093/aje/kwx348.

Peskoe, S. B. *et al.* (2021) 'Adjusting for selection bias due to missing data in electronic health records-based research', *Statistical Methods in Medical Research*. SAGE Publications Ltd, 30(10), pp. 2221–2238. doi: 10.1177/09622802211027601.

Pierce, C. A. *et al.* (2008) 'Surveillance bias and deep vein thrombosis in the national trauma data bank: The more we look, the more we find', *Journal of Trauma - Injury, Infection and Critical Care*. J Trauma, 64(4), pp. 932–936. doi: 10.1097/TA.0b013e318166b808.

Ponsford, M. J. *et al.* (2020) 'Clinical and laboratory characteristics of clozapine-treated patients with schizophrenia referred to a national immunodeficiency clinic reveals a B-cell signature resembling common variable immunodeficiency (CVID)', *Journal of Clinical Pathology*. BMJ Publishing Group, 73(9), pp. 587–592. doi: 10.1136/jclinpath-2019-206235.

Rajagopal, S. (2005) 'Clozapine, agranulocytosis, and benign ethnic neutropenia', *Postgraduate Medical Journal*. BMJ Publishing Group, 81(959), pp. 545–546. doi: 10.1136/pgmj.2004.031161.

Rodová, A. *et al.* (1973) 'A blind comparison of clozapine and perphenazine in schizophrenics.', *Activitas Nervosa Superior*. Act Nerv Super (Praha), 15(2), pp. 94–95.

Ross, M. K., Wei, W. and Ohno-Machado, L. (2014) '"Big data" and the electronic health

record.', *Yearbook of medical informatics*. Thieme Medical Publishers, 9(1), pp. 97–104. doi: 10.15265/IY-2014-0003.

Rössler, W. *et al.* (2005) 'Size of burden of schizophrenia and psychotic disorders', *European Neuropsychopharmacology*. Eur Neuropsychopharmacol, 15(4), pp. 399–409. doi: 10.1016/j.euroneuro.2005.04.009.

Van Rossum, J. M. *et al.* (1970) 'Pharmacology', *Modern problems of pharmacopsychiatry*, 5, pp. 23–70.

Seaman, S. R. and White, I. R. (2013) 'Review of inverse probability weighting for dealing with missing data', *Statistical methods in medical research*. Stat Methods Med Res, 22(3), pp. 278–295. doi: 10.1177/0962280210395740.

Seeman, P. (2014) 'Clozapine, a fast-off-D2 antipsychotic', *ACS Chemical Neuroscience*. American Chemical Society, 5(1), pp. 24–29. doi: 10.1021/cn400189s.

Shoenfeld, Y. *et al.* (1988) 'Benign familial leukopenia and neutropenia in different ethnic groups', *European Journal of Haematology*. Eur J Haematol, 41(3), pp. 273–277. doi: 10.1111/j.1600-0609.1988.tb01192.x.

Siskind, D., Honer, W. G., *et al.* (2020) 'Consensus statement on the use of clozapine during the COVID-19 pandemic', *Journal of psychiatry & neuroscience : JPN*. NLM (Medline), 45(3), p. 2. doi: 10.1503/jpn.200061.

Siskind, D., Sidhu, A., *et al.* (2020) 'Systematic review and meta-analysis of rates of clozapine-associated myocarditis and cardiomyopathy', *Australian and New Zealand Journal of Psychiatry*. SAGE PublicationsSage UK: London, England, 54(5), pp. 467–481. doi: 10.1177/0004867419898760.

Siskind, D. *et al.* (2021) 'Rates of treatment-resistant schizophrenia from first-episode cohorts:

systematic review and meta-analysis', *The British Journal of Psychiatry*. Cambridge University Press, pp. 1–6. doi: 10.1192/bjp.2021.61.

Stephens, P. (1990) 'A review of clozapine: An antipsychotic for treatment-resistant schizophrenia', *Comprehensive Psychiatry*. W.B. Saunders, 31(4), pp. 315–326. doi: 10.1016/0010-440X(90)90038-T.

Stewart, R. *et al.* (2009) 'The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data', *BMC Psychiatry*, 9(1), p. 51. doi: 10.1186/1471-244X-9-51.

Stewart, R. (2014) 'The big case register', *Acta Psychiatrica Scandinavica*. Acta Psychiatr Scand, 130(2), pp. 83–86. doi: 10.1111/acps.12279.

Stoecker, Z. R. *et al.* (2017) 'Clozapine usage increases the incidence of pneumonia compared with risperidone and the general population: A retrospective comparison of clozapine, risperidone, and the general population in a single hospital over 25 months', *International Clinical Psychopharmacology*. Lippincott Williams and Wilkins, 32(3), pp. 155–160. doi: 10.1097/YIC.0000000000000162.

Stroup, T. S. *et al.* (2016) 'Comparative effectiveness of clozapine and standard antipsychotic treatment in adults with schizophrenia', *American Journal of Psychiatry*. Am J Psychiatry, 173(2), pp. 166–173. doi: 10.1176/appi.ajp.2015.15030332.

Stroup, T. S. and Gray, N. (2018) 'Management of common adverse effects of antipsychotic medications', *World Psychiatry*. World Psychiatric Association, 17(3), pp. 341–356. doi: 10.1002/wps.20567.

Tate, A. R. *et al.* (2009) 'Determining the date of diagnosis - Is it a simple matter? the impact of different approaches to dating diagnosis on estimates of delayed care for ovarian cancer in UK

primary care', *BMC Medical Research Methodology*. BioMed Central, 9(1), pp. 1–9. doi: 10.1186/1471-2288-9-42.

Tate, A. R. *et al.* (2011) 'Using free text information to explore how and when GPs code a diagnosis of ovarian cancer: An observational study using primary care records of patients with ovarian cancer', *BMJ Open*. British Medical Journal Publishing Group, 1(1), p. e000025. doi: 10.1136/bmjopen-2010-000025.

Taylor, D., Barnes, T. R. E. and Young, A. H. (2018) *The Maudsley prescribing guidelines in psychiatry*. London: Wiley-Blackwell.

Taylor, D. M. *et al.* (2012) 'Augmentation of clozapine with a second antipsychotic - a meta-analysis', *Acta Psychiatrica Scandinavica*. Acta Psychiatr Scand, 125(1), pp. 15–24. doi: 10.1111/j.1600-0447.2011.01792.x.

Ting, E. *et al.* (2019) 'Does the frequency of administration of long acting injectable antipsychotics impact psychiatric outcomes and adverse effects: A systematic review and meta-analysis', *Journal of Psychiatric Research*. J Psychiatr Res, 109, pp. 193–201. doi: 10.1016/j.jpsychires.2018.12.004.

Torniainen, M. *et al.* (2015) 'Antipsychotic treatment and mortality in schizophrenia', *Schizophrenia Bulletin*. Schizophr Bull, 41(3), pp. 656–663. doi: 10.1093/schbul/sbu164.

Wahlbeck, K. *et al.* (1999) 'Evidence of clozapine's effectiveness in schizophrenia: a systematic review and meta-analysis of randomized trials.', *The American journal of psychiatry*, 156(7), pp. 990–9. doi: 10.1176/ajp.156.7.990.

Walker, A. M. *et al.* (1997) 'Mortality in current and former users of clozapine', *Epidemiology (Cambridge, Mass.)*. Epidemiology, 8(6), pp. 671–677. doi: 10.1097/00001648-199710000-00010.

Weinberger, D. R. and Harrison, P. J. (2011) 'Schizophrenia, Third Edition', *Schizophrenia: Third*

Edition. Wiley-Blackwell. doi: 10.1002/9781444327298.

Weinmann, S., Read, J. and Aderhold, V. (2009) 'Influence of antipsychotics on mortality in schizophrenia: Systematic review', *Schizophrenia Research*. Schizophr Res, 113(1), pp. 1–11. doi: 10.1016/j.schres.2009.05.018.

WHO (1992) *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. World Health Organization.

World Health Organization (1995) *Physical status: the use and interpretation of anthropometry. Report of a WHO Expert Committee.*, *World Health Organization technical report series*. Switzerland. doi: 10.1002/(sici)1520-6300(1996)8:6<786::aid-ajhb11>3.0.co;2-i.

Xiao, C., Choi, E. and Sun, J. (2018) 'Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review', *Journal of the American Medical Informatics Association*. Oxford University Press, 25(10), pp. 1419–1428. doi: 10.1093/jamia/ocy068.

Zhao, J. and Henriksson, A. (2016) 'Learning temporal weights of clinical events using variable importance', *BMC Medical Informatics and Decision Making*. BioMed Central Ltd, 16. doi: 10.1186/s12911-016-0311-6.

ZTAS (2018) *Zaponex Treatment Access System (ZTAS) manual*. Nijmegen, Netherlands: Leyden Delta.